Klassifikation der Rentenversicherten anhand der Verläufe von Entgeltpunkten

Classification of those contributing to the German pension fund on the basis of contributory points earned

BACHELORARBEIT

ZUR ERLANGUNG DES AKADEMISCHEN GRADES BACHELOR OF SCIENCE (B.Sc)

IN VOLKSWIRTSCHAFTSLEHRE

AN DER WIRTSCHAFTSWISSENSCHAFTLICHEN FAKULTÄT DER

HUMBOLDT-UNIVERSITÄT ZU BERLIN

VORGELEGT VON

IVAN MITKOV

MATRIKEL-NR.: 542007

Prüfer: Prof. Dr. W. Härdle

Betreuer: Dr. S. Klinke

BERLIN, 24. MÄRZ 2014

Danksagung

Ich möchte an dieser Stelle Herrn Dr. Sigbert Klinke für die Betreuung und für die Motivation im Bereich der Statistik danksagen.

Ich bedanke mich an meine Eltern, dass sie mein Studium in Deutschland verwirklicht haben.

Ich danke meiner Freundin, dass sie mich immer unterstützt und inspiriert.

INHALTSVERZEICHNIS

ABBILDUNGSVERZEICHNIS

TABELLENVERZEICHNIS

1 Einleitung

Die Wurzeln der deutschen Rentenversicherung verbergen sich im Gesetz zu der Invaliditäts- und Alterssicherung, das Bismarck 1889 verabschiedet hat. In den vergangenen 125 Jahre hat Deutschland verschiedene politische Wandel ausgehalten, schwere Kriegszeiten überwunden, eine gnadenlose Hyperinflation besiegt, Teilung und Wiedervereinigung erlebt. Jedes dieser Ereignisse hat seinen Beitrag für die Gestaltung der Rentenversicherung über die Jahreszeiten hinweg.

Die gegenwärtige Zeit hat auch ihre eigene Besonderheiten, die die Rentenversicherung prägen. Im 21. Jahrhundert ist die Medizin schon sehr fortgeschritten. Dafür sprechen sowohl die erfolgreiche Bekämpfung unheilbarer Krankheiten aus der nächsten Vergangenheit als auch die gesunkene Kindersterblichkeit. Die Globalisierung der Welt verursacht den schnellsten technologischen Fortschritt aller Zeiten. An dieser Situation versuchen die Menschen sich besser anzupassen, indem sie immer mehr in die Ausbildung und die Verbesserung der eigenen Fähigkeiten investieren. Aus diesem Grund werden der Arbeitsmarkteintritt und die Gedanken an Nachwuchs auf spätere Jahre verschoben. Alle dargestellten Fakten sind grundlegend für den demographischen Wandel im 21. Jahrhundert, der anhand in der Abbildung 1.1 dargestellten Veränderungen der Bevölkerungspyramide von 1910 bis 2050 sichtbar gemacht wird.

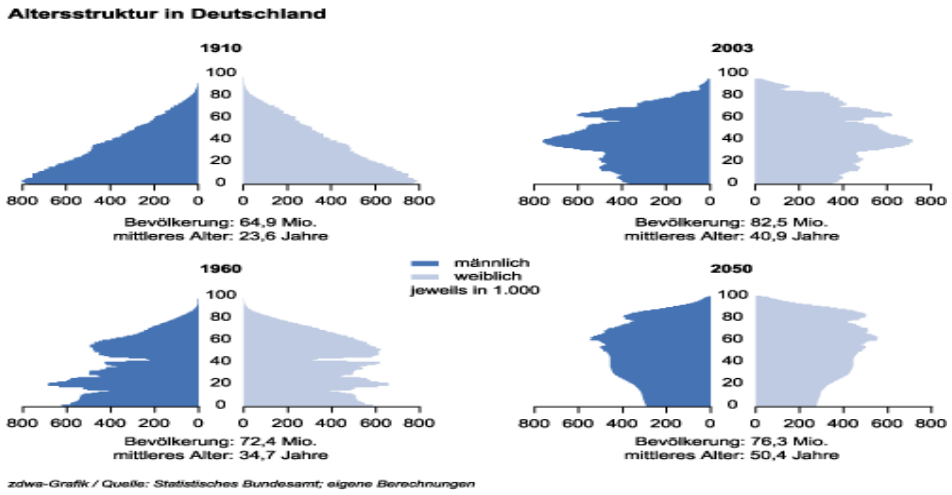


Abbildung 1: *Altersstruktur in Deutschland*

Falls die Regierung ein stabiles Rentensystem für die kommenden Generationen sichern möchte, sollte der Staat dem demographischen Wandel und den Problemen, die daraus resultieren, durch verschiedene Maßnahmen im Bereich der Rentenversicherung entgegenwirken.

So kommt man zum Jahr 1989, in dem die erste Reform eingeführt wurde. Sie strebt eine stufenweise Erhöhung der Altersgrenze für Männer und Frauen an. Anders gesagt, die Reform beeinflusst durch die Verschiebung der oberen Grenze des Erwerbsalters auf indirekter Weise das Verhalten der Menschen auf dem Arbeitsmarkt. Für diesen eindeutigen Zusammenhang zwischen der Erwerbsdauer, dem verdienten Einkommen und den zukünftigen Renten fallen die Entgeltpunkte ins Gewicht. Dieser Begriff wird im folgenden Abschnitt näher betrachtet, weil er zentral für die Analyse ist.

1.1 Was ist ein Entgeltpunkt?

Der Entgeltpunkt ist ein Maß für die Quantifizierung des Zusammenhangs zwischen dem persönlichen Einkommen während des Erwerbsalters und der Rente. Anders ausgedrückt entspricht die Anzahl der verdienten Entgeltpunkte eines Jahres genau dem Verhältnis zwischen dem persönlichen Einkommen und dem Durchschnittseinkommen für dasselbe Jahr, wobei das Durchschnittseinkommen von Anlage 1 SGB VI abzulesen ist. Die Formel für einen Entgeltpunkt (EP) sieht so aus:

$$EP = \frac{Eigenes\ beitragpflichtiges\ Einkommen}{Durchschnittseinkommen\ nach\ Anlage\ 1}$$

Beispiel: Es wird angenommen, dass die berücksichtigte Person für das Jahr *X* 50 000 Euro verdient hat, während das Durchschnittseinkommen für dieses Jahr 32 446 Euro beträgt. Daraus folgt:

$$\frac{50\,000}{32\,446} = 1{,}54\ Entgeltpunkte$$

Die versicherte Person hat für das Jahr *X* 1,54 Entgeltpunkte erworben.

Damit die Differenz im Einkommen zwischen den Ostdeutschen und den Westdeutschen abgemildert wird, kommt der Hochrechnungsfaktor zunutze. Er wird jedes Jahr neu berechnet und auch

vom Gesetzbuch bestimmt. Mit dieser Kennzahl wird das Einkommen der Versicherten, die in Ostdeutschland arbeiten, multipliziert. Somit ergeben sich für einen Hochrechnungsfaktor von 1,1740 die folgende Anzahl von Entgeltpunkten:

$$\frac{1{,}1740 * 50\,000}{32\,446} = 1{,}81\ Entgeltpunkte$$

Man soll auch die Beitragsbemessungsgrenze, die wieder von der Regierung festgelegt wird, berücksichtigen. Als Beitragsbemessungsgrenze wird das maximale Bruttoeinkommen bestimmt, bis zu welchem die maximale Beitrage zur Rentenversicherung errechnet werden. Mit anderen Worten bekommt man nach der Überschreitung dieser Grenze keine Erhöhung der Entgeltpunkte mehr.

Schließlich sind die berechneten Entgeltpunkte für alle Jahre des Erwerbsalters aufzusummieren und mit dem aktuellen Rentenwert eines Entgeltpunktes zu multiplizieren.

Beispiel: Die berücksichtigte Person aus Ostdeutschland geht schon in die Rente. Sie hat insgesamt 30 Entgeltpunkte erworben, wobei der aktuelle Rentenwert 25 Euro beträgt. Dann würde sie eine monatliche Rente von 750 (30*25=750) Euro erhalten.

1.2 Ziel der Arbeit

Der Fokus dieser Arbeit liegt auf der Klassifikation der deutschen Rentenversicherten anhand der Verläufe ihrer Entgeltpunkte. Dabei werden die Entgeltpunkte nach den Kalendermonaten aufgeteilt analysiert. Das zielt auf dem Zusammenhang zwischen den Entgeltpunkten und den Jahreszeiten. Die monatliche Verteilung der Entgeltpunkte wird zunächst anhand drei der wichtigsten Merkmale einer Verteilung berechnet und zwar des Mittelwerts, des Medianes und der Standardabweichung aller erworbenen Punkte im Monat Januar. Analog werden dann die drei Kennzahlen auch für die restlichen Monate ermittelt. Als Nächstes wird mit allen neuen 36 Variablen (12 Monate je 3 neueingeführten Variablen für Mittelwert, Median und Standardabweichung) ein globales Clustering durchgeführt, mit dessen Hilfe die Menschen nach ihren Verläufen in Gruppen unterteilt werden und zwar so, dass Personen mit ähnlichen Verläufen einer Gruppe zugeordnet werden. Zur statistischen Absicherung der Ergebnisse wird das ganze Instrumentarium der Clusteranalyse im

Softwareprogram SPSS, angewendet. Diese Herangehensweise dient zur Absicherung der Ergebnisse, weil jedes Verfahren zu einigen Ungenauigkeiten führen könnte. Nach der Gruppierung der Rentenversicherten werden die einzelnen Cluster näher betrachtet, um festzustellen, welche Personen in jeder Klasse von Verläufen enthalten sind. Für die Interpretation der Ergebnisse werden schließlich einige soziodemographischen Merkmale und Werte aus der Rentenberechnung herangezogen, da es sich zeigen wird, dass eine erneute Durchführung der Clusteranalyse unter deren Berücksichtigung, unter bestimmten Bedingungen, die Klassifikation der Rentenversicherten verbessern kann.

2 Clusteranalyse

Die Klassifikation der Verläufe der Entgeltpunkte erfolgt anhand der Clusteranalyse. Somit wird das Verfahren grundsätzlich für die Arbeit und fordert deshalb einen näheren Blick. Bevor man jedoch zu der Datenanalyse kommt, sind unterschiedliche Algorithmen des Clusterings in Einzelheiten kennenzulernen. In diesem Abschnitt werden die theoretischen Aspekte, die Vor- und Nachteile jedes Verfahrens des Clusterings besprochen.

2.1 Definition

Die Clusteranalyse hat zum Ziel eine Gruppierung von Objekten, wobei die Objekte durch unterschiedliche Merkmale charakterisiert werden. Das Cluster-Algorithmus soll die Versicherten nach der Ähnlichkeit ihrer Eigenschaften klassifizieren, wobei die gebildeten Gruppen intern möglichst homogen und extern möglichst unterschiedlich sein sollen. Am Ende der Analyse soll jedes Objekt genau zu einem Cluster zugeordnet werden. „Entscheidend für das Ergebnis einer Clusteranalyse ist die Definition der Ähnlichkeit von Objekten bzw. Clustern und die Art des Optimierungskriteriums, mit dem man eine möglichst gute Separation der Cluster erzielen will“.[1] Nur, wenn das

[1] Bortz, Jürgen: Statistik für Human- und Sozialwissenschaftler, Springer, 6. Auflage, Berlin, 2004

Distanzmaß ausgewählt wurde, kann man zum nächsten Schritt, der Ausführung der Clusteranalyse, weitergehen. Aus diesem Grund werden die Ähnlichkeits- und Distanzmaßen in einem separaten Abschnitt berücksichtigt.

2.2 Ähnlichkeits- und Distanzmaße

Damit das Ähnlichkeitsmaß ermittelt wird, ist das Vergleichen von Objekten nach ihren Merkmalen notwendig. Da die Skalierung dieser Merkmale eine entscheidende Rolle spielt, werden im Folgenden die üblichsten Methoden zur Ermittlung der Ähnlichkeits- und Distanzmaße in Abhängigkeit vom Skalenniveau dargestellt, wobei für die Clusteranalyse in der vorliegenden Arbeit die Kennzahlen für intervallskalierten Variablen verwendet werden.

2.2.1 Nominalskalierte Merkmale

Es gibt eine Reihe von Möglichkeiten, das Ähnlichkeitsmaß zweier nominalskalierten Merkmale zu berechnen. Da SPSS nur dichotome Merkmale erkennt, werden nur sie im Detail besprochen.

Im Fall binärer Variablen sind vier Möglichkeiten zu berücksichtigen. Ein Beispiel stellt Tabelle 1 dar, wobei die Ausprägung 1 eine Vorhandensein des Merkmals und die Ausprägung 0 eine Nichtvorhandensein des Merkmals bedeuten.

		Objekt 1	
		1	0
Objekt 2	1	a	b
	0	c	d

Tabelle 1: *4-Felder-Tabelle zur Ermittlung des Ähnlichkeitsmaßes*

Darin steht *a* für die Anzahl der Merkmale, die bei den beiden Versicherten mit 1 ausgeprägt sind, *b* steht für 1 von der Seite von Objekt 2 und 0 von der Seite von Objekt 1, *c* bedeutet 0 für Person 2 und 1 für Person 1 und schließlich repräsentiert *d* die Situation, wenn die beiden Objekten eine

Ausprägung von 0 haben. Für die Ermittlung der Ähnlichkeits- und Distanzmaße bei dichotomen Variablen werden üblicherweise die folgenden Koeffizienten verwendet:

- Der S-Koeffizient wurde von Jaccard bzw. Rogers und Tanimoto entwickelt. Er betont die Bedeutung des gleichzeitigen Vorhandenseins an dem Anteil aller Ausprägungen, die für mindestens einen der Versicherten eine 1 enthalten. Die Formel sieht folgendermaßen aus

$$S_{ij} = \frac{\mathrm{a}}{\mathrm{a} + \mathrm{b} + \mathrm{c}}$$

  während das Distanzmaß lautet

$$d_{ij} = \frac{b + c}{a + b + c}$$

- Der SMC-Koeffizient oder Simple-Matching-Koeffizient schließt noch das gleichzeitige Fehlen eines Merkmals ein. Die modifizierte Gleichung ist:

$$SMC_{ij} = \frac{a + d}{a + b + c + d}$$

  Und das dazugehörige Distanzmaß: *1-SMC_{ij}*.

- Die Phi-Koeffizient gibt allen Feldern das gleiche Gewicht, wobei die Distanz durch *1-Φ* dargestellt wird. Man muss auf die Randverteilung aufpassen, weil Φ davon abhängig ist.

Weitere Ähnlichkeitsmaße werden in Tabelle 2 dargestellt.

Würfel	$2a/(2a + b + c)$
Russel und Rao	$a/(a + b + c + d)$
Kulczynski	$a/(b + c)$
Yule	$(ad - bc)/(ad + bc)$
Ochiai	$\sqrt{\frac{a}{a+b}\frac{a}{a+c}}$

Tabelle 2: *Ähnlichkeitsmaße für binäre nominalskalierte Variablen*

K-Fach gestufte Merkmale betreffen die nominalskalierten Merkmale, die mehr als zwei Ausprägungen haben. Mit Hilfe von Dummy-Variablen werden zusätzliche Variablen erstellt, wobei jede neue Variable genau einer Kategorie entspricht. Es werden insgesamt k – 1 Variablen zu dem Datensatz hinzufügt. Mit diesem Ähnlichkeitsmaß sollte man vorsichtig vorgehen, weil nominale Merkmale mit den meisten Kategorien am schwersten gewichtet werden. Das kann durch die Gewichtung jeder Ausprägung der Merkmale mit 1/(k – 1) vermieden werden.

2.2.2 Ordinalskalierte Merkmale

Für ordinalskalierte Merkmale gibt es mehrere Möglichkeiten, die Ähnlichkeit zu ermitteln. Hier werden die üblichsten Verfahren vorgestellt, die allerdings nicht unproblematisch sind

- Rangplatzierung ist eine Methode, die jeder Ausprägung der Variable eine Zahl angibt, wobei die höchste Kategorie der größten Zahl entspricht und die am niedrigsten geordnete Kategorie die kleinste Zahl bekommt. Hier kann das Problem entstehen, dass die Ränge als Ausprägungen einer intervallskalierten Variable behandelt werden und eine genaue Aussage über den Abstand zwischen den Rangplätzen sich nicht machen lässt. Aus diesem Grund wird es an dieser Stelle empfohlen, die einzelnen Kategorien der Variable zu dichotomisieren.
- Kendalls τ ist ein Maß für die Korrelation von ordinalskalierten Variablen, mit dessen Hilfe auch die Ähnlichkeit bestimmt werden kann. Die Kennzahl vergleicht jedes Paar der Ausprägungen einer Variable mit dem entsprechenden Paar der anderen Variable. Es lassen sich demzufolge die folgenden fünf Kategorien definieren:.
    - $x_i < x_j$, $y_i < y_j$, deren Anzahl ist C.
    - $x_i < x_j$,$y_i > y_j$, deren Anzahl ist D.
    - $x_i \neq x_j$, $y_i = y_j$, deren Anzahl ist T_Y.
    - $x_i = x_j$, $y_i \neq y_j$, deren Anzahl ist T_X.
    - $x_i = x_j$, $y_i = y_j$, deren Anzahl ist T_{XY}.

    Mit Hilfe dieser Gleichungen und Ungleichungen ergibt sich die endgültige Formel:

$$\frac{C - D}{\sqrt{(C + D + T_X)(C + D + T_Y)}}$$

2.2.3 Intervallskalierte Merkmale

Auf diesen Abschnitt ist eine besondere Rücksicht zu nehmen, weil im Rahmen dieser Arbeit eine Clusteranalyse mit Hilfe der Entgeltpunkte durchgeführt wird. Die Entgeltpunkte sind offensichtlich intervallskaliert, deswegen sind die unten dargestellten Abstandsmaßen der Ausgangspunkt für die weiteren Analysen.

Im Falle der intervallskalierten Variablen spricht man nicht mehr über ein Ähnlichkeitsmaß, sondern über das Distanzmaß. Im Folgenden werden die bekanntesten Distanzmaße bei intervallskalierten Variablen beschrieben:

- Die Euklidische Metrik oder auch Euklidischer Abstand ist das Maß, das am häufigsten für die Ermittlung der Distanz zweier Objekte gebraucht wird. Die Kennzahl ergibt sich von dem Quadratwurzel der Summe der quadrierten Abstände:

$$\sqrt{\sum_{j=1}^{p}(X_j - Y_j)^2}$$

  Ein mögliches Problem, das rechtzeitig zu berücksichtigen ist, stellen die unterschiedlichen Maßstäbe der Merkmale der Objekte dar. So wird es beispielsweise sinnlos sein, die Distanz zwischen X_j und Y_j zu berechnen, wenn die Zahlen von X_j für Prozente stehen, während diese für Y_j die Temperaturen Celsius angeben. Ein Lösungsweg, der in solchen Situationen häufig angewendet wird, ist die Z-Transformation[2].

- Die Quadrierte Euklidische Distanz ist fundamental für die Ausführung von der hierarchischen Clusteranalyse mit SPSS. „Die quadrierte Euklidische Distanz errechnet sich damit als die Summe der quadrierten Differenzen zwischen den Variablenwerten der beiden zu vergleichenden Objekte."[3]

$$\sum_{j=1}^{p}(X_j - Y_j)^2$$

  Wegen des großen Rechenaufwands hat dieses Maß allerdings den Nachteil, dass es nur für kleine-Stichproben einsetzbar ist.

- Tschebysheff gibt als Distanzmaß den größten Abstand eines Wertepaares an:

$$Tschebyschev\ Distanz = \max_i |X_i + Y_i|$$

- City-Block-Distanz oder Mannhatan Metrik stellt die summierten Differenzen aller Wertepaare dar:

[2]Unter Z-Transformation ist die Standardisierung einer Variable zu verstehen, wobei der Erwartungswert Null ist und die Varianz den Wert von Eins hat.

[3]Brosius, Felix: SPSS 16: Das mitp-Standardwerk, Fundierte Einführung in SPSS und die Statistik, mitp, 1. Auflage, 2008

$$Block\ Distanz = \sum |X_i + Y_i|$$

- Minkowski-Distanz drückt sich in der Generalisation von der Euklidischen Distanz und der Manhattan Distanz aus:

$$Minkowski\ Distanz = \sqrt[p]{\sum (X_i + Y_i)^p}$$

wobei für *p* ein Wert zwischen 0 und 1 auszuwählen ist.

2.2.4 Maße für Häufigkeiten

Für Clusteranalyse mit Variablen, die absoluten Häufigkeiten repräsentieren, stehen im Softwareprogramm SPSS zwei Maße zur Verfügung:

- Das Chi-Quadrat-Maß summiert die quadrierten Abweichungen der beobachteten mit dem Wert der erwarteten Häufigkeit auf. Für die Ermittlung des Maßes für zwei Fälle mit *m* Variablen sieht die Formel wie folgt aus:

$$\chi^2 = \sum_{i=1}^{2} \sum_{j=1}^{m} \frac{(n_{ij} - n_{ij}^*)^2}{n_{ij}^*}$$

- Das Phi-Quadrat-Maß wird für die Ermittlung der Ähnlichkeit zwischen zwei dichotomen Variablen angewendet. Ausgangspunkt für die Errechnung des Koeffizients ist die Tabelle vom Abschnitt 2.2.1.
  Daraus ergibt sich:

$$\Phi = \frac{ad - bc}{\sqrt{(a+b)(a+c)(d+c)(d+b)}}$$

2.3 Verfahren

In clusteranalytischen Verfahren werden die einzelnen Cluster, nach Bühl & Zöfel (2000), durch schrittweise Fusionierung gebildet. Zentralle Rolle spielen die hierarchischen und partitionierenden Verfahren, wobei letztere insbesondere bei großen Fallzahlen angewendet werden. Da die vor-

liegende Stichprobe relativ groß ist, werden alle dieser Arten der Clusteranalyse, die im SPSS implementiert sind, zunutze kommen. Jede der Methoden ist für gewisse Umstände geeignet. Das erfordert eine detaillierte Betrachtung, damit jede Umsetzung der Verfahren begründbar wird.

2.3.1 Hierarchische Clusteranalyse

Die hierarchische Clusteranalyse ist ein der grundlegenden Verfahren des Clusterings. Es werden zwei Typen von Verfahren unterschieden. Einerseits steht die *agglomerative Clusteranalyse*, bei der jeder Versicherte am Anfang als einzelner Cluster betrachtet wird, andererseits beruht das *divise Verfahren* auf die Tatsache, dass alle Objekte zu Beginn der Analyse sich in einer Klasse befinden und später unterteilt werden.

Im Folgenden ist das agglomerative Verfahren zu beachten, weil es das Relevante für die Arbeit ist. Nachdem jeder Versicherte als ein Cluster repräsentiert wird, muss man die benachbarten Cluster vereinigen. Mit anderen Worten werden diese Objekte, die nach ihren Merkmalen am ähnlichsten sind, fusioniert. Die Vereinigung der Cluster kann so lange fortgesetzt werden bis zum Zeitpunkt, wenn alle Objekte in einer Gruppe zusammengefasst werden. Hier ist jedoch Vorsicht geboten, weil in den letzteren Schritten ein Zusammenschließen zwischen weit voneinander stehenden Clustern möglich ist. Damit das vermieden wird, wird die Distanz immer in Betracht eingezogen. Die Vereinigung von Clustern muss dann abgebrochen werden, wenn die Erhöhung des Abstandsmaßes stark von den bis zu diesem Moment betrachteten abweicht. Im Allgemeinen ist der Verlauf der hierarchischen Clusteranalyse ähnlich der Struktur eines Baumes, wie das Dendrogramm in Abbildung 2 zeigt. Im Dendrogramm werden auch die Distanzen zwischen den Clustern mittels der vertikalen Kanten dargestellt.

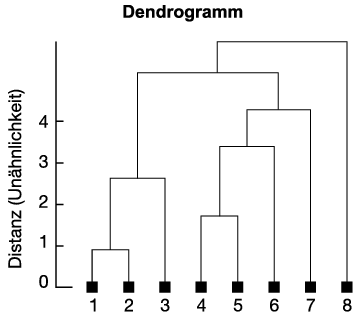


Abbildung 2: *Dedrogramm*

Für die Fusionierung zweier Cluster werden verschiedene Methoden verwendet. Die Einzelheiten der relevantesten werden im Folgenden erläutert.

- *Linkage zwischen den Gruppen* ist eine Art, bei der die Distanzen aller möglichen Paare von Objekten zu ermitteln sind. Die Paare werden so konstruiert, dass aus jedem Cluster je ein Objekt ausgewählt wird. Am Ende wird als Abstandsmaß zwischen den Clustern das arithmetische Mittel aller Distanzen genommen.
- *Linkage innerhalb der Gruppen* unterscheidet sich von der oben genannten Methode dadurch, dass die Paare auch Objekte desselben Clusters enthalten können. Die Technik der Berechnung bleibt aber unverändert. Das Abstandsmaß repräsentiert das arithmetische Mittel der Distanzen aller Paare.
- *Nächststehender Nachbar* ist eine Methode, bei der aus jedem Cluster nur ein Objekt zu beachten ist. Die Objekte werden so ausgewählt, dass ihre Distanz möglichst kleiner ist. Die Cluster, deren Paar den kleinsten Abstand hat, werden vereinigt.
- *Entferntester Nachbar* ist nach der größten Distanz zweier Objekte, die zu unterschiedlichen Clustern gehören, ausgerichtet. Es werden diejenige Cluster fusioniert, deren maximale Differenz am kleinsten ist.
- Für die *Zentroid Clustering* werden die Variablenmittelwerte aller Fälle im Cluster berechnet. Das Abstandsmaß wird wie Distanz zweier Variablen erhalten mit dem einzigen Unterschied, dass die Variablenwerte durch die Mittelwerte ersetzt werden.
- In der *Ward Methode* werden die Mittelwerte aller Variablen im Cluster berechnet. Danach werden die quadrierten euklidischen Distanzen der Objekte zum Clustermittelwert ermittelt und miteinander aufsummiert. Schließlich werden diejenige Cluster vereinigt, deren Fusionierung den kleinsten Zuwachs für die summierten Abstände hat.

Nachdem die Wesentlichkeit der hierarchischen Clusteranalyse dargestellt wurde, kann man auch die entsprechenden Schlussfolgerungen über dieses Verfahren ziehen. Die hierarchische Clusteranalyse hat in Bezug auf die Feststellung der Clusteranzahl einen klaren Vorteil gegenüber den anderen Arten von Clustering. Wie beschrieben, kann man die Clusteranzahl selbst anhand der Abstandskoeffizienten bestimmen. Die hierarchische Clusteranalyse verfügt ebenso über vielfältige Abstands- und Distanzmaßen, womit das Verfahren bei allen Skalierungen einsetzbar ist. Zu den Schwächen der hierarchischen Clusteranalyse ist zu erwähnen, dass die Abstandskoeffizienten

ohne Vorinformation über Clusterzentren und die Anzahl der Cluster großen Rechenaufwand verursachen. Um das zu vermeiden, kann man sich nur auf kleinere Stichproben begrenzen. Ein weiterer negativer Aspekt der hierarchischen Clusteranalyse ist, dass die Zuordnung eines Objekts zu einem Cluster unrevidierbar ist, deswegen sollte der Output mit diesem von anderen Cluster-Verfahren verglichen werden. Aus diesem Grund wird in der vorliegenden Arbeit auf die Clusterzentrenanalyse und auf die Two-Step Clusteranalyse zurückgegriffen. Die beiden Verfahren ermöglichen die Analyse der vollständigen Stichprobe, dass zu genaueren Aussagen über die Klassifikation der Rentenversicherten führt.

2.3.2 Clusterzentrenanalyse

Dieses Verfahren gehört zu der *partitionierenden Clusteranalyse*, bei der man über Vorinformation über die Gruppen (die Cluster) und ihre Anzahl verfügen soll. Ist das nicht der Fall, besteht die Möglichkeit, die Anfangswerte mit SPSS zu ermitteln. In den weiteren Schritten werden die Objekte von Cluster zu Cluster verschoben, wobei jedes Mal das Objekt zum nächsten Clusterzentrum zugeordnet wird. Diese Vorgehensweise hat zum Ziel, die interne Homogenität und die externe Differenzierbarkeit der Klassen zu verbessern.

Die Schritte der Clusterzentrenanalyse werden im Folgenden beschrieben. Zu Beginn ist das Vorliegen einer Vorinformation, die am besten mit dem hierarchischen Verfahren zu erhalten ist, von Vorteil. Dafür braucht man oftmals eine Unterstichprobe, damit der Rechenaufwand vermieden wird. Man erhält daraus eine erste grundsätzliche Anzahl der Klassen und Aufteilung der Objekte, wobei für jeden Cluster das Clusterzentrum berechnet wird. Als Clusterzentrum wird der Vektor bezeichnet, der die Mittelwerte der einzelnen Variablen für jeden Cluster repräsentiert. Nach der Einbeziehung der restlichen Fälle wird jedes Objekt zu dem nächsten Cluster zugeordnet. Hier kommt die von Kapitel 2.2.3 bekannte quadrierte euklidische Distanz zunutze:

$$d(i,j) = \left\|X_i - C_j\right\|^2 = \sum_{q=1}^{Q} (x_{qi} - c_{qj})^2$$

wobei:

X_i → steht für den Vektor des i-ten Falls

C_j → steht für das Clusterzentrum des Clusters J

Q → Anzahl der Variablen

$x_{qi}\rightarrow$ gibt den Wert der q-ten Variable von dem i-ten Fall an

$c_{qj} \rightarrow$ gibt den Wert der q-ten Variable von dem j-ten Fall an

Nach der ersten Sortierung aller Fälle werden wieder die Clusterzentren ermittelt und die Fälle nochmal zu dem nächsten Cluster verschoben. Dieser Prozess dauert so lange, bis die Maximalzahl der Iterationen oder der Wert vom Konvergenzkriterium erreicht wird.

Schließlich lässt sich das Verfahren wie folgt beurteilen: Die Clusterzentrenanalyse ist eine Methode, die aufgrund der Vereinfachung der Berechnungen, für große Stichproben geeignet ist. Das hat allerdings den Nachteil, dass für die Durchführung der Analyse eine Vorinformation vorausgesetzt wird. Da diese Auskunft durch das hierarchische Verfahren oder von SPSS selbst ermittelt wird, führen die Variation der Unterstichproben beziehungsweise die Vielfalt von den mit SPSS ermittelten Anfangswerten zu großen Verzerrungen. Des Weiteren kann die partitionierende Clusteranalyse möglicherweise zu einem Cluster keine weiteren Objekte verschieben und auf dieser Weise leere Klassen bilden. Ein weiterer Kritikpunkt bezüglich des Verfahrens ist die Reihenfolge der Fälle. Wird der Datensatz umsortiert, so können auch neue Anfangswerte entstehen.

2.3.3 Two-Step Clusteranalyse

Ein anderes Clusterverfahren, das in die Arbeit eingesetzt wird, ist die Two-Step Clusteranalyse. „Der von SPSS verwendete Algorithmus der zweistufigen Clusteranalyse wurde in dieser Form von der Firma SPSS selbst entwickelt, basiert jedoch auf einem gängigen Verfahren, dem so genannten BIRCH-Algorithmus (Balanced Iterative Reducing and Clustering using Hierarchies). Dieser Algorithmus wurde von SPSS insbesondere um die Möglichkeit zur Verarbeitung kategorialer Daten erweitert.“[4] Einerseits ist die Durchführung der Two-Step Clusteranalyse notwendig, um die Ergebnisse der vorigen Verfahren zu bestätigen , andererseits dient die Two-Step Clusteranalyse zur Schätzung der Qualität des Clusterings anhand des Silhouettenkoeffizients, der am Ende dieses Kapitels ausführlich erläutert wird.

Wie es eingeführt wurde, ist das Verfahren für diskrete und/oder stetige Variablen geeignet. Für diesen Zweck stehen zwei Distanzmaßen zur Verfügung. Das Log-Likelihood-Maß findet genau

[4]Brosius, Felix: SPSS 16: Das mitp-Standardwerk, Fundierte Einführung in SPSS und die Statistik, mitp, 1. Auflage, 2008

dann eine Anwendung, falls die einbezogenen Merkmale sowohl kategorial als auch stetig sind, während das schon bekannte Euklidische Distanzmaß nur für stetige Variablen vernünftig ist.

Im Folgenden wird die Log-Likelihood-Distanz kurz erläutert. Sie stellt die Ähnlichkeit zweier Objekte oder Cluster anhand einer Variable dar, wobei ihr Mittelwert und ihre Varianz berechnet werden. Angenommen teilt die Clusteranalyse die Fälle in zwei Gruppen auf, so werden für jede davon wiederum der Mittelwert und die Varianz ermittelt. „Geht man nun davon aus, dass die Variable in der Grundgesamtheit einer bestimmten Verteilung folgt, so lässt sich die Wahrscheinlichkeit ausrechnen, mit der in einer Stichprobe der tatsächlich beobachtete Mittelwert sowie die tatsächlich beobachtete Varianz auftreten, wenn die beiden Cluster(Fallgruppen) in ihren Eigenschaften (Mittelwert, Varianz, Größe) repräsentativ für die Grundgesamtheit sind“.[5] Entscheidend hierbei ist die Beachtung der Verteilung von Objekten unter den Klassen. Die Formel für die Log-Likelihood Distanz der beiden Cluster *i* und *j* lautet:

$$d(i,j) = \xi_i + \xi_j - \xi_{<i,j>}$$

wobei

$$\xi_v = -N_v\left(\sum_{k=1}^{K^A}\frac{1}{2}log\left(\hat{\sigma}_k^2 + \hat{\sigma}_{vk}^2\right) + \sum_{k=1}^{K^B}\hat{E}_{vk}\right)$$

und

$$\hat{\mathrm{E}}_{vk} = -\sum_{l=1}^{L_k}\frac{N_{vkl}}{N_v}log\frac{N_{vkl}}{N_v}$$

Hier ist zu ergänzen, was die Notation bedeutet.

K^A → Anzahl der metrischen Variablen

K^B → Anzahl der kategorialen Variablen

L_k → Anzahl der Kategorien von der k-ten kategorialen Variable

N → Anzahl aller Fälle

N_v → Anzahl der Fälle des Clusters *v*

$\hat{\sigma}_k^2$→ geschätzte Varianz von der k-ten metrischen Variable

$\hat{\sigma}_{vk}^2$ →geschätzte Varianz von der k-ten metrischen Variable des Clusters *v*

[5]Brosius, Felix: *SPSS 16: Das mitp-Standardwerk*, *Fundierte Einführung in SPSS und die Statistik*, mitp, 1. Auflage, 2008

N_{vkl}→Anzahl der Fälle von Cluster v, deren k-te kategoriale Variable die l-te Kategorie annimmt

Die Log-Likelihood Distanz hat aber auch einige Voraussetzungen, die entscheidend für ihre Anwendung sind. Die stetigen Variablen müssen normalverteilt sein, während die diskreten der multinominalen Verteilung folgen müssen. Alle Variablen müssen außerdem noch unabhängig voneinander sein, was durch Kreuztabellen, Korrelationskoeffizienten oder geeignete Tests überprüft werden kann.

Wie es sich vom Namen des Verfahrens erkennen lässt, wird die Two-Step Clusteranalyse mittels zwei Schritte durchgeführt.

In der *ersten Stufe* wird jeder Fall einzeln betrachtet und wird entschieden, ob der Fall zu einem existierenden Cluster zugeordnet werden muss oder eine neue Gruppe zu gründen ist. Für dieses Ziel werden die Fälle des Datensatzes in einer baumartigen Struktur umgeformt. So wird das CF-Tree aufgebaut, wobei die oberste Ebene dem Ursprung (der Wurzel) des Baumes entspricht. Die Gesamtheit der Objekte wird von der Wurzel repräsentiert und die unterste Ebene des Baumes - als Blätter dargestellt. Bevor die Analyse jedoch durchgeführt wird, ist die genaue Anzahl der Ebenen der Baumstruktur zu bestimmen. Im Endergebnis ist jedes Objekt genau einer Klasse zugeordnet.

Nach der ersten Unterteilung aller Fälle anhand eines der Distanzmaße werden anhand der geeigneten Ähnlichkeits- und Distanzmaße mehrere Sub-Cluster oder die sogenannten Äste gebildet. Im nächsten Schritt werden von jeder Fallgruppe neue Klassen (die Knoten des Baumes) gegründet. Die Unterteilung wird auf dieser Weise bis zu der untersten Ebene (die Blätter) fortgesetzt. Da werden die Fälle zu einem des bestehenden Clusters in Rahmen jedes einzelnen Blattes zugeordnet, wobei der vorbestimmten Schwellenwert für die Heterogenität jedes Clusters nicht überschritten werden darf. Sollte dies nicht der Fall sein, muss ein neuer Cluster gegründet werden.

Das Hauptziel der ersten Stufe beruht auf der Reduzierung des Rechenaufwands für große Stichprobenumfänge und das Endergebnis wird als der Ausgangspunkt für die *zweite Stufe* verwendet. Als Nächstes müssen die von den Blättern gebildeten Sub-Cluster miteinander fusioniert werden. Der Software, mit dem die Arbeit ausgewertet wurde, benutzt die schon dargestellte hierarchische Clusteranalyse für die Vereinigung der Gruppen. Wie es in Kapitel 2.3.2 erläutert wurde, kennt man in diesem Fall die Anzahl der Cluster nicht. Sie lassen sich automatisch ermitteln, wobei zwei

Clusterkriterien zur Verfügung stehen: das Akaikes-Informationskriterium (AIC) und Bayes-Informationskriterium (BIC). Jedes davon legt fest, wie die Clusteranazahl berechnet wird. Für den Fall mit J Clustern, lassen sich AIC und BIC so darstellen:

$$BIC\ (J) = -2\sum_{j=1}^{J} \xi_j + m_J \log(N)$$

$$AIC\ (J) = -2\sum_{j=1}^{J} \xi_j + 2m_j$$

wobei

$$m_j = J\left\{2K^A + \sum_{k=1}^{K^B}(L_k - 1)\right\}$$

Es wurde eingeleitet, dass die Ausgabe von der Two-Step Clusteranalyse in SPSS den Silhouettenkoeffizient enthält. Die Kennzahl repräsentiert die Qualität der Zuordnung eines Objektes zu den beiden Clustern in nächster Nähe. In der unterstehenden Formel wird der Silhouettenkoeffizient eines Clusters *C* gezeigt:

$$S_C = \frac{1}{n_C}\sum_{o \in C} S(o)$$

wobei

$$S(o) = 0 \text{ , falls } dist(A, o) = 0 \text{ gilt}$$

$$S(o) = \frac{dist(B, o) - dist(A, o)}{max\{dist(A, o), dist(B, o)\}}$$

dist(A, o) → gibt die Distanz des Objekts *o* zu dem Cluster A an

dist(B, o) → gibt die Distanz des Objekts *o* zu dem Cluster B an

Mit anderen Worten wird dabei das arithmetische Mittel aller Silhouetten innerhalb des Clusters berechnet. Die Kennzahl S_C kann Werte im Intervall [0;1] annehmen, wobei die Werte in drei Kategorien aufgeteilt sind. / $0{,}00 < S_C \leq 0{,}25$ für schlechte Cluster-Qualität; $0{,}25 < S_C \leq 0{,}5$ für mittlere Cluster-Qualität und $0{,}5 < S_C \leq 1$ für gute Cluster-Qualität/

Schließlich soll das Verfahren auch beurteilt werden. Die Two-Step Clusteranalyse hat den Vorzug, dass es auch bei riesengroßen Datensätzen angewendet werden kann. Ins Verfahren können außerdem gleichzeitig stetige und diskrete Variablen einbezogen werden. Nicht zu unterschätzen ist, dass die Güte des Clusterings von dem Silhouettenkoeffizient abgelesen werden kann. Man sollte aber bei der ersten Stufe vorsichtig sein, weil die Struktur des Cluster Feature Baumes von der Reihenfolge der Fälle im Datensatz beeinflusst werden kann. Aus diesem Grund kann möglicherweise eine Unstabilität des Clusterings vorkommen. Um mit diesem Problem umgehen zu können, soll man die Homogenität in den Ausgaben des Verfahrens überprüfen.

3 Anwendung der Clusteranalyse

Für die Analysen in dieser Arbeit werden die Daten des Forschungsdatenzentrums der Deutschen Rentenversicherung verwendet. Der Datensatz mit dem Namen „PUFVSKT2007“ gehört zu den Public Use Files für die Lehre und kann von allen Studierenden bestellt werden. Die Daten wurden als eine geschichtete Zufallsstichprobe der deutschen Rentenversicherten im Jahr 1983 erhoben. Seitdem wird der Datensatz als Panel gepflegt und fortgeführt. Der Public Use File wurde 2007 veröffentlicht, wobei die gezogene Stichprobe 5% der deutschen Rentenversicherten von den Jahrgängen von 1940 bis einschließlich 1977 repräsentiert. Die Stichprobe umfasst 13 359 Menschen und enthält Information über ihre soziodemographischen Merkmale und ihre biographiebezogenen Verlaufsmerkmale.

3.1 Die Methoden

Der Datensatz des Forschungsdatenzentrums ist für die Nutzung von SPSS aufbereitet. Für die Clusteranalysen werden alle Verfahren, die in der Software implementiert sind, eingesetzt. Dabei

wird auch die Two-Step-Clusteranalyse, deren Algorithmus von den Entwicklern von SPSS erstellt wurde, zunutze kommen. Für die Abbildung der Grafiken wird auch Excel angewendet.

3.2 Deskriptive Statistik

Damit der Inhalt des Datensatzes für den Leser deutlicher wird, werden im Folgenden die für die Arbeit wichtigsten Variablen vorgestellt.

Für die Schlussfolgerungen und die Interpretation der Outputs werden die Ergebnisse immer auf dem Hintergrund der soziodemographischen Merkmale betrachtet, wobei die Outputs zunächst nach dem Geschlecht verglichen werden. Das Verhältnis zwischen Männern und Frauen sieht folgendermaßen aus:

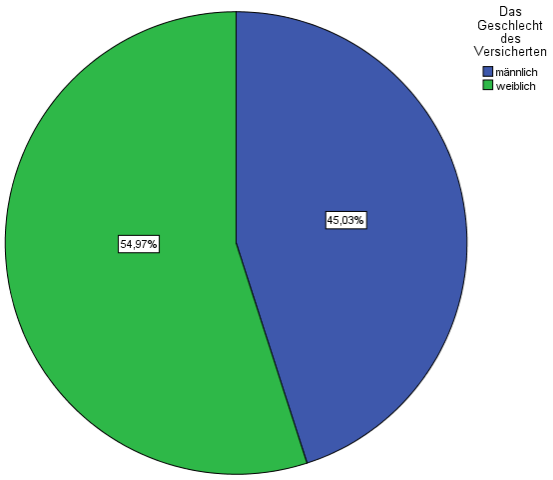


Abbildung 3: *Geschlechterverhältnis im Datensatz. (Männer - grün; Frauen - blau)*

Des Weitern soll noch ein Blick auf die Zeiträume, in denen die Versicherten geboren sind, geworfen werden. Als Unterstützung der These für die Relevanz des Alters ist hervorzuheben, dass neben den Änderungen der Entgeltpunkte aufgrund des Alters diese Variable auch verschiedene historische Ereignisse bzw. Zeiten repräsentieren könnte. Das Alter der Menschen kann zum Beispiel die Periode vor und nach dem politischen Wandel von 1989 oder die Zeiten, als eine Geburt mit siebzehn für etwas Normales angenommen wird, umfassen. Damit diese möglichen Auffälligkeiten in den Verläufen der Versicherten erklärt werden und die Interpretation der Cluster verfeinert wird,

ist das Geburtsjahr der Untersuchungspersonen zu berücksichtigen. Das Balkendiagramm dieser Variable veranschaulicht, dass Daten aus allen Altersgruppen zur Verfügung stehen. (s. Abb. 4)

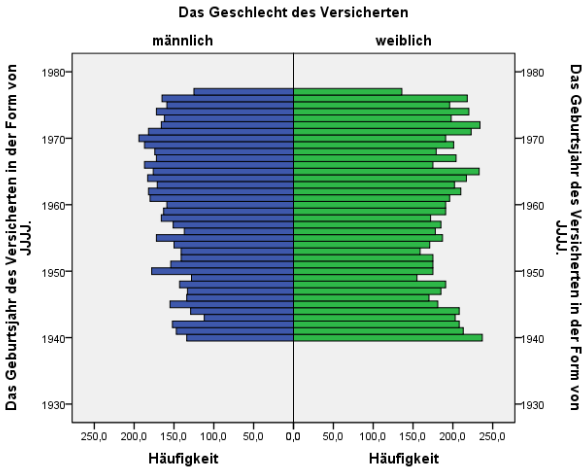


Abbildung 4: *Histogramm des Alters, aufgeteilt nach dem Geschlecht*

Im Folgenden werden die Histogramme einiger Werte aus der Rentenberechnung dargestellt. Die Merkmale werden für die Interpretation der Cluster umgesetzt.

Als vollwertige Beitragszeiten werden die Monate bezeichnet, die ausschließlich Beitragszeiten enthalten. Die Anzahl dieser Monate hängt direkt vom Alter des Versicherten ab.

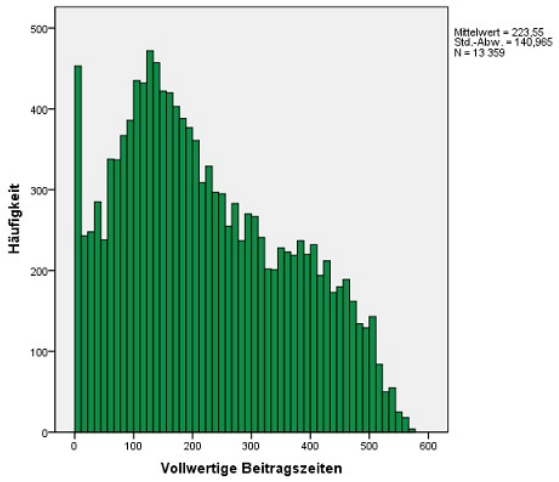


Abbildung 5: *Histogramm der vollwertigen Beitragszeiten*

Die Abbildung 5 veranschaulicht, dass keine der Personen für alle 624 Monate ausschließlich Beitragszeiten gehabt hat. Grund dafür sind die beitragsgeminderten Zeiten und die Anrechnungszeiten, wobei als beitragsgemindert diese Monate bezeichnet werden, die sowohl Beitragszeiten als

auch beitragsfreie Zeiten aufweisen. Als Anrechnungszeiten werden in der Deutschen Rentenversicherung die Monate gekennzeichnet, für die keinen Beitrag gezahlt wurde (vgl. Abb. 6).

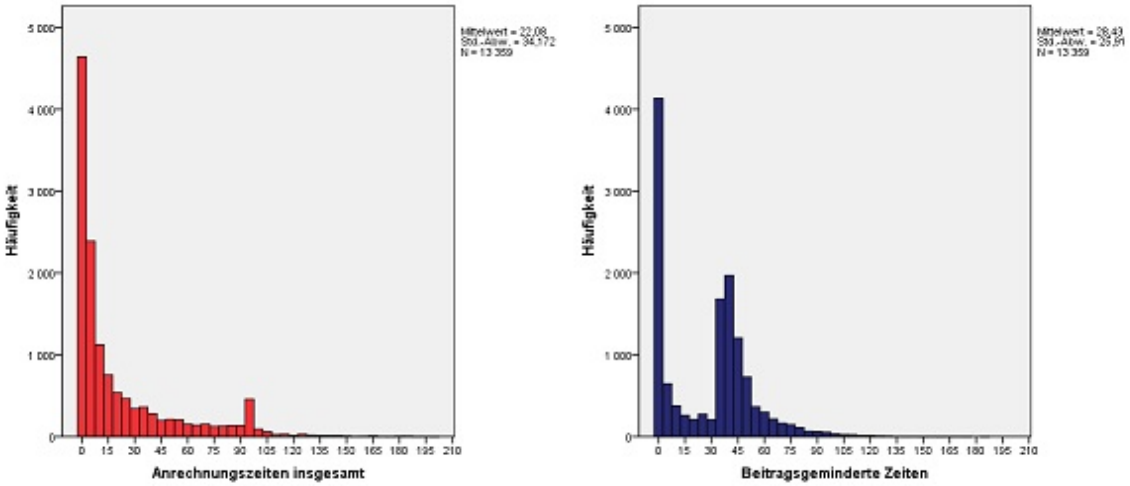


Abbildung 6: *Histogramme aller Anrechnungszeiten und der beitragsgeminderte Zeiten*

3.3 Globale Clusteranalyse

Als globale Clusteranalyse wird im Rahmen der Arbeit die Clusteranalyse angegeben, die alle Versicherte von jedem Alter einbezieht. Mit Hilfe dieser Analyse werden die Menschen grundsätzlich klassiert, wobei alle Verfahren angewendet werden. Im nächsten Schritt werden die daraus kommenden Outputs miteinander verglichen. Im Abschluss werden die einzelnen Klassen anhand der soziodemographischen Merkmale und der Werte aus der Rentenberechnung charakterisiert.

3.3.1 Die Variablen

Wie es schon eingeführt ist, hat die vorliegende Arbeit zum Ziel, die Rentenversicherten nach ihren Verläufen von Entgeltpunkten zu klassifizieren. Aus diesem Grund sind die MEGPT_*i* Variablen im Datensatz für die weiteren Analysen entscheidend. Die Variablen basieren auf der sozialen Erwerbssituation der Versicherten und repräsentieren die verdienten Entgeltpunkte für jeden Monat. Hier ist zu beachten, dass die Anhebungen wegen der Kindererziehung sowie die Punkte für geringes Einkommen nicht enthalten sind.

Die Variable MEGPT_i soll sorgfältiger betrachtet werden. Der Index i im Namen der Variablen kommt vor, weil er für den Monat der Variable steht Der Betrag von Eins im Index steht für Januar für das Jahr, in dem die untersuchte Person 14 Jahre alt geworden ist. Die letzten Variablen von den biographiebezogenen Merkmalen ist MEGPT_624, was Dezember 52 Jahre später ist. So kann durch Addition von 12 leicht berechnet werden, dass MEGPT_1, MEGPT_13, MEGPT_25 usw. Januar, während MEGPT_2, MEGPT_14, MEGPT_26 Februar darstellen. Analog findet man die entsprechenden Variablen für jeden Monat im Kalenderjahr.

Die Verteilung der Entgeltpunkte über die Jahre wurde bereits beschrieben. Das in Kombination mit der vorgestellten Technik für die Bestimmung des Kalendermonats, wird in den kommenden Clusteranalysen zu mehr Genauigkeit beitragen. Die Monatsaufteilung ist für die Identifizierung und eventueller Klassifikation in einzelnen Gruppen der Saisonarbeiter sehr gut geeignet. Es werden drei der für eine Verteilung aussagekräftigsten Kennzahlen eingeführt, damit die angestrebte Clusteranalyse realisiert wird. Das sind nämlich der Mittelwert, der Median und die Standardabweichung. Während der Mittelwert und der Median Maße der zentralen Tendenz sind und den Durchschnittswert (für intervallskalierten Variablen) bzw. den häufigsten Wert (für ordinalskalierte Variablen) darstellen, gibt die Standardabweichung (nur bei intervallskalierten Daten) zusätzlich Auskunft über die Streuung der Einzelwerte um den Mittelwert sowie über mögliche Ausreißer, die die Verteilung stark beeinflussen können. Für jeden der zwölf Monate werden diese drei Kennzahlen berechnet und dadurch 36 neuen Variablen zum Datensatz hinzugefügt.

Die für das globale Clustering relevantesten Variablen, die eine zusätzliche Erläuterung benötigen, werden im Folgenden ausführlich vorgestellt. Die unterstehende Tabelle 3 fasst alle Variablen, die in die Analyse einbezogen sind, mit einer kurzen Beschreibung zusammen

Variable	Erläuterung
MEGPT_i	Gibt die verdienten Entgeltpunkte für jeden einzelnen Monat der Vesicherten an
MEAN_j	Der Mittelwert der verdienten Entgeltpunkte für alle Monate, die j sind, wobei j für den Namen des entsprechenden Kalendermonats steht.
MEDIAN_j	Der Median der verdienten Entgeltpunkte für alle Monate, die j sind, wobei j für den Namen des entsprechenden Kalendermonats steht.
SD_j	Die Standardabweichung der verdienten Entgeltpunkte für alle Monate, die j sind, wobei j für den Namen des entsprechenden Kalendermonats steht.
GEH	Geschlecht der Untersuchungsperson
GBJA	Geburtsjahr des Versicherten in der Form JJJJ
GBKIJ1	Geburtsjahr des ersten Kindes
GBKIJ2	Geburtsjahr des zweiten Kindes
GBKIJ3	Geburtsjahr des dritten Kindes
BYVL	Vollwertige Beitragszeiten
BYGM	Beitragsgeminderte Zeiten
AZ	Anrechnungszeiten insgesamt
AUAZ	Anrechnungszeiten wegen Krankheit
AJAZ	Anrechnungszeiten wegen Arbeitslosigkeit
SCHULAZ	Summe der Anrechnungszeiten wegen schulischer Ausbildung
Alter	Alter der Untersuchungsperson
Alter_1Kind	Alter der Frau bei der ersten Geburt
Alter_2Kind	Alter der Frau bei der zweiten Geburt
Alter_3Kind	Alter der Frau bei der dritten Geburt

Tabelle 3: *Erläuterung der Variablen von der globalen Clusteranalyse*

3.3.2 Ergebnisse der hierarchischen Clusteranalyse

Die theoretischen Ansätze der hierarchischen Clusteranalyse hat man detailliert im Kapitel 2.3.1 dargelegt. In diesem Abschnitt wird das Verfahren anhand des Datensatzes durchgeführt. Da das hierarchische Cluster-Verfahren von großem Rechenaufwand begleitet ist, kann man in diesem Fall eine Zufallsstichprobe von allen Beobachtungen im Datensatz ziehen. Dabei wird die Erhaltung von Vorinformation für die geeignete für große Datensätze k-Means-Analyse angestrebt.

Bei der Ziehung der Unterstichproben ist zu beachten, dass die unterschiedlichen Stichproben unterschiedliche Ergebnisse zur Folge haben können. Aus diesem Grund wurde mehrmals eine Zufallsstichprobe von 1500 Beobachtungen gezogen. Für Anzahl der Cluster wurde diese Anzahl angenommen, die von den einzelnen hierarchischen Analysen am häufigsten auftritt.

Nach der Ziehung der benutzten Unterstichprobe, ergibt sich die folgende Häufigkeitstabelle (siehe Tabelle4).

Verarbeitete Fälle[a]

Fälle					
Gültig		Fehlenden Werten		Insgesamt	
N	Prozent	N	Prozent	N	Prozent
1461	97,4%	39	2,6%	1500	100,0%

a. Quadrierte Euklidische Distanz wurde verwendet

Tabelle 4: *Tabelle der verarbeiteten Fälle von der hierarchischen Clusteranalyse*

Die in der Tabelle enthaltenen fehlenden Werte, signalisieren, dass in diesem Fall weniger Vereinbarungsschritte notwendig sind. Zur Erinnerung wird hier nochmal erwähnt dass die Fusionierung der Sub-Cluster dann abgebrochen werden muss, wenn ein beschleunigter Anstieg des Abstandskoeffizentes zu beobachten ist.

Da man durch die Clusteranalyse die Rentenversicherten klassieren möchte, soll man sich in SPSS für eine Gruppierung der Fälle entscheiden. Bevor es aber zu der echten hierarchischen Clusteranalyse kommt, müssen die Cluster-Methode und das Distanzmaß ausgewählt werden. „Da einige dieser Methoden (Nächstgelegener Nachbar, Entferntester Nachbar) offensichtliche Nachteile haben, andere nur noch schwer durchzuschauen sind, ist es wohl zu empfehlen, die voreingestellte und einsichtige Methode »Linkage zwischen den Gruppen« zu verwenden.“[6] Die Variablen, die

[6]Bühl, Achim; Zöfel, Peter: SPSS Version 10, Einführung in die moderne Datenanalyse unter Windows, ADISSON-WESLEY, 7., überarbeitete und erweiterte Auflage, München, 2000

die monatlichen Mittelwerte, Mediane und Standardabweichungen repräsentieren, sind intervallskaliert, deswegen nimmt man auch die voreingestellte quadrierte euklidische Distanz. Zusammengefasst sehen die Einstellungen wie folgt aus:

- Cluster: nach Fällen klassieren
- Cluster-Methode: Linkage zwischen den Gruppen
- Distanzmaß: Quadrierter Euklidischer Abstand
- Standardisieren: Ja, mit Z-Werten

Zuordnungsübersicht

	Zusammengeführte Cluster			Erstes Vorkommen des Clusters		
Schritt	Cluster 1	Cluster 2	Koeffizienten	Cluster 1	Cluster 2	Nächster Schritt
1	23	710	,000	0	0	12
...	...	...	...	...	...	...
1455	3	20	89,277	1454	1450	1457
1456	29	283	95,836	1452	1440	1458
1457	3	701	111,225	1455	0	1459
1458	1	29	116,423	1449	1456	1459
1459	1	3	193,660	1458	1457	1460
1460	1	691	461,101	1459	0	0

Tabelle 5: *Zuordnungsübersicht der hierarchischen Clusteranalyse*

In Tabelle 5 werden die einzelnen Schritte der Vereinigung der Cluster(im Fall entspricht jeder Cluster genau einer Person) dargestellt. Berücksichtigt man die Koeffizienten des Abstands zwischen den fusionierten Gruppen, ist der Koeffizient des 1459 Schrittes von besonderer Bedeutung. Der Wert von 193,660 hat einen deutlich größeren Anstieg, bezogen zu dem Anstieg der vorigen Koeffizienten, deswegen muss man bei Schritt 1458 die Vereinbarung der Cluster abbrechen. Also hier ist die nächste Fusionierung der Sub-Cluster in zwei Sub-Cluster zu vermeiden. Das bedeutet, dass die Menschen der gezogenen Zufallsstichprobe am besten in drei Cluster verteilt werden können.[7]

[7]Es werden separat auch andere Stichproben gezogen, die fast in allen Fällen das gleiche Ergebnis haben. Aus diesem Grund wird angenommen, dass die benutzte Stichprobe repräsentativ für den Datensatz ist.

3.3.3 Ergebnisse der Clusterzentrenanalyse

Damit alle Rentenversicherten in die Analyse eingeschlossen werden, wendet man die Clusterzentrenanalyse[8] an. Im Kapitel 2.3.2 wurde erläutert, dass das Verfahren eine Vorinformation benötigt, aufgrund dessen die aus dem hierarchischen Clustering erhaltenen Zentren als Anfangswerte gelten. Normalerweise werden die Variablen für diese Methode auch standardisiert, aber im Rahmen der Arbeit wird eine Ausnahme gemacht. Das zielt die Beibehaltung der Punkte im Originalform, was trotz der kleineren Werte zu einer einfacheren Interpretation beiträgt.

Von der hierarchischen Analyse wird die optimale Anzahl der Cluster ermittelt, wobei die Mittelwerte der Variablen für die einzelnen Gruppen von der Tabelle der Clusterzentren der endgültigen Lösung abzulesen sind (siehe Anhang A.1).

Nach der Einbeziehung aller Versicherten wird die k-Means Clusteranalyse mit den folgenden Voreinstellungen durchgeführt:

- Clusteranazahl: 3 (nachgewiesen mit dem hierarchischen verfahren)
- Maximale Interrationen: 10
- Konvergenzkriterium: 0
- Anfangswerte: diese von A.1 im Anhang anwenden.
- Es wird auch eine Variable über die Zugehörigkeit der einzelnen Personen zu einem Cluster generiert. Die Variable ist erforderlich, damit die Ergebnisse dieses Verfahrens mit diesen von der Two-Step Clusteranalyse verglichen werden

Eine neue Sortierung der Untersuchungspersonen ist in diesem Fall nicht erforderlich, da die anfänglichen Werte von den anhand der hierarchischen Clusteranalyse errechneten Zentren übernommen werden.

Nach der Ausführung des Clusterings ergeben sich die folgenden Ergebnisse:

[8]Jedes Mal, wenn die Clusterzentrenanalyse in Betracht kommt, wird die k-Means-Methode gedacht.

Iteration	Änderung in Clusterzentren		
	1	2	3
1	,004	,002	,002
2	,000	,000	,000
3	,000	,000	,000
4	,000	,000	,000
5	5,278E-5	7,262E-5	7,091E-5
6	,000	1,954E-5	1,513E-5
7	,000	2,050E-5	1,586E-5
8	,000	2,084E-5	1,613E-5
9	,000	,000	,000

Distanz zwischen Clusterzentren der endgültigen Lösung

Cluster	1	2	3
1		,400	,219
2	,400		,184
3	,219	,184	

Tabelle 6: *Änderung in Clusterzentren bei den einzelnen Iterationen (links); Distanz zwischen Clusterzentren der endgültigen Lösung (rechts)*

Von Interesse ist hier das Iterationsprotokoll, weil die Zuordnung der Objekte nach dem neunten Schritt abgebrochen wird. Grund dafür ist die zu kleine Änderung der Clusterzentren. Von der rechten Tabelle ist die Distanz zwischen den Clustern zu sehen, wobei der kleinste Abstand 0,184 beträgt. Nach dem Ende des Verfahrens werden neue Vektoren für die Clusterzentren erhalten. Die neuen Werte findet man in Tabelle A.2 im Anhang.

Der Output enthält weitere Information über die Verteilung der Rentenversicherten über die einzelnen Clustern. Von der ausgegebenen Tabelle ist leicht erkennbar, dass die Häufigkeit der Untersuchungspersonen mit den größten monatlichen Zuwächsen der Entgeltpunkte am niedrigsten ist.

Anzahl der Fälle in jedem Cluster

Cluster	1	2117,000
	2	4793,000
	3	6193,000
Gültig		13103,000
Fehlend		256,000

Tabelle 7: *Anzahl der Fälle in jedem Cluster nach der Einbeziehung aller Versicherten*

Als Nächstes werden die Klassen anhand der Angaben aus der Tabelle mit den endgültigen Lösungen beschrieben. Damit die Schätzung robuster gegen Ausreißer in den Beobachtungen ist, wird als üblicher monatlicher Zuwachs der Medianwert der einzelnen Monate angenommen.

- Cluster 1 vereinigt 2 117 Personen. In diesem Fall haben die Monate mit 31 Tagen einen Zuwachs von 0,12 Entgeltpunkten, während die anderen einen Anstieg von 0,11 aufweisen.[9] Allgemein kann hierzu gesagt werden, dass diese Menschen die größten Zuwächse haben.
- Cluster 2 enthält 4 793 Rentenversicherten, wobei die Personen die niedrigsten monatlichen Zuwächse von 0,03 haben
- Cluster 3 umfasst 6 193 Menschen. Diese Personen haben die mittleren Zuwächse von 0,07.

Die Klassen werden in den folgenden Abschnitten ausführlich charakterisiert, weil eine Bestätigung des Clusterings erstmal notwendig ist.

3.3.4 Überprüfung des Clusterings anhand der Two-Step Clusteranalyse

Damit die erhaltenen Ergebnisse an Plausibilität gewinnen, wird für die globale Klassierung der Fälle auch das Two-Step Verfahren eingesetzt. Die Methode bringt Nutzen für die Analyse, weil sie neben der erneuten automatischen Ermittlung der Cluster auch den Silhouetenkoeffizient ausgibt. Außerdem kann man alle Rentenversicherten in die Analyse einschließen, ohne eine Vorinformation zu verwenden.

Im ersten Schritt werden die gleichen Variablen ausgewählt, wobei sie auch zum Standardisieren angegeben wurden. Da die Entgeltpunkte stetig sind, entscheidet man sich für das Log-Likelihood-Distanzmaß. Die ausgewählten Einstellungen sehen folgendermaßen aus:

- Nur stetige Variable eingeführt
- Distanzmaß: Log-Likelihood[10]
- Anzahl der Cluster: automatisch (maximal 15)
- Cluster-Kriterium: BIC

[9] Durch die unterschiedliche Dauer der Monate (30/31 Tage) ergeben sich bei voll belegten Monaten geringfügige Schwankungen (im Kalendermonat Februar fallen diese etwas größer aus).

[10] Da die Variablen metrisch sind, kann auch die Euklidische Distanz verwendet werden, wobei aber großer Rechenaufwand entsteht.

- Speichern: Es wird auch, wie in der Clusterzentrenanalyse, eine Variable über die Zugehörigkeit der Versicherten zu einer Klasse.

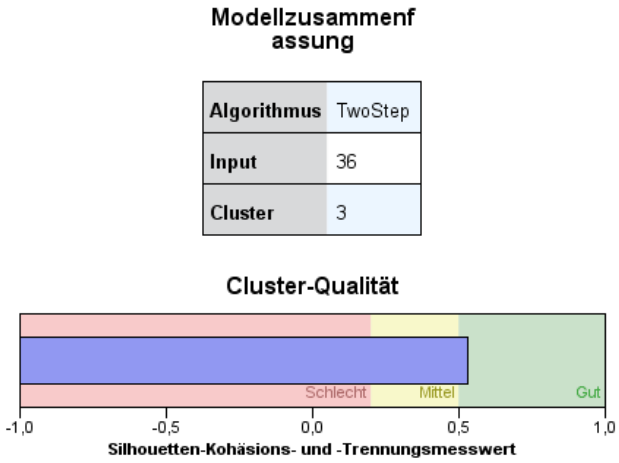


Abbildung 7: *Zusammenfassung und Cluster-Qualität des Two-Step Clusterings*

Vom dargestellten Output wird nachvollziehbar, dass dieses Verfahren die Rentenversicherten auch in 3 Gruppen klassiert. Der ausgegebene Silhouettenkoeffizient sagt aus, dass die Aufteilung zwischen den Klassen gut ist. Es wird separat die Stabilität der Two-Step Clusteranalyse geprüft. Für diesen Zweck werden die Fälle mehrmals neu sortiert, es werden auch einige Stichproben aus dem Datensatz gezogen, die den gleichen Output aufweisen.

Damit die Gleichheit der Ergebnisse aus der Clusterzentrenanalyse und dem Two-Step Algorithmus statistisch bestätigt wird, wendet man die Kreuztabelle an. Diese Methode dient als statistisches Instrument zur Veranschaulichung der Zuordnung der Personen zu den unterschiedlichen Klassen.

Da die generierten Variablen für die Zugehörigkeit einer Person zu einem Cluster die Ausprägungen „1“, „2“, „3“ haben, wobei die Zahlen keine Ordnung der Cluster repräsentieren, kann man schließen, dass die Variablen kategorialskaliert sind.

Eine Methode, die die Verteilung der Untersuchungspersonen zwischen den Kategorien der beiden Variablen darstellen kann, ist die Kreuztabelle.

			Zugehörigkeit der Versicherten zu einem Cluster nach der Two-Step Clusteranalyse			
			1	2	3	Gesamt
Zugehörigkeit der Versicherten zu einem Cluster nach der Clusterzentrenanalyse	1	Anzahl	0	334	1783	2117
		Erwartete Anzahl	853,4	926,3	337,3	2117,0
	2	Anzahl	507	5399	289	6195
		Erwartete Anzahl	2497,3	2710,5	987,2	6195,0
	3	Anzahl	4775	0	16	4791
		Erwartete Anzahl	1931,3	2096,2	763,5	4791,0
Gesamt		Anzahl	5282	5733	2088	13103
		Erwartete Anzahl	5282,0	5733,0	2088,0	13103,0

Tabelle 8: *Kreuztabelle der Clusterzugehörigkeit*

Von den Werten der Tabelle 8 ist leicht zu erkennen, dass die Versicherten fast auf der gleichen Weise klassiert sind. Dafür spricht die große Häufigkeit der Verteilung in bestimmten Kategorien. Hier soll allerdings berücksichtigt werden, dass die gleichen Ausprägungen der beiden Variablen nicht für die gleichen Cluster stehen. Von der Abbildung 8 des 3-D Balkendiagramms ist es offensichtlich, dass Cluster 3 der Clusterzentrenanalyse dem Cluster 1 von der Two-Step Clusteranalyse entspricht.

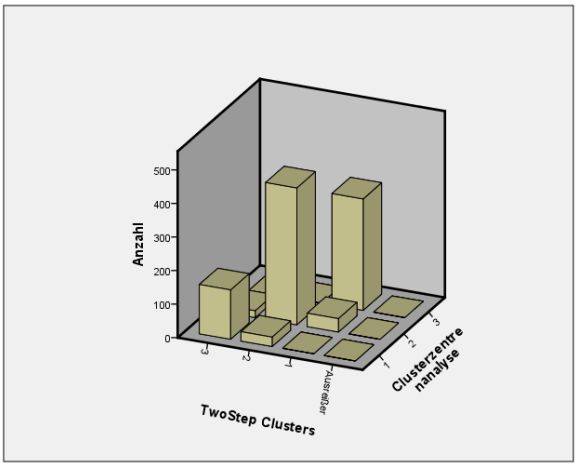


Abbildung 8: *3-D Balkendiagramm der Clusterzugehörigkeit*

Nach diesen Überprüfungen, ob die Clusterzentrenanalyse und das Two-Step Verfahren die gleiche Anzahl der Gruppen und die gleichen Cluster ausgeben, kann zusammengefasst werden, dass die

erhaltene Klassifikation der deutschen Rentenversicherten mittels der Clusterzentrenanalyse glaubhaft ist. Aus diesem Grund kann man mit der Interpretation der Cluster fortführen.

3.3.5 Interpretation des globalen Clusterings

In diesem Abschnitt werden die einzelnen Klassen näher betrachtet. Das zielt auf die Beantwortung der Frage, von welchen Menschen die Cluster besetzt sind. Ausgangspunkt für die Interpretation der Gruppen ist die Clusterzentrenanalyse, wobei die Clusterzentren schon ermittelt wurden. Die drei Cluster wurden schon grob charakterisiert (siehe Kapitel 3.3.3), wobei die monatlichen Zuwächse jeder einzelnen Gruppe anhand der monatlichen Mediane aus dem Vektor der Clusterzentren bestimmt werden. Cluster 1 wird als die Klasse der Menschen mit großen Zuwächsen festgelegt, wobei der übliche Versicherte im Schnitt 0,12-0,13 Entgeltpunkte pro Monat verdient. Cluster 2 ist von Menschen, die die kleinsten Zuwächse (etwa 0,03 monatlich) aufweisen, besetzt. Cluster 3 enthält die Personen mit den Zuwächsen von 0,06 oder mit anderen Worten sind das die Versicherten, die durchschnittlich verdienen.

Um die Interpretation der Cluster verfeinern zu können, werden die soziodemographischen Merkmale zum ersten Mal in die Analyse einbezogen. Als Erstes wird anhand der Variable „GBJA“ (Geburtsjahr des Versicherten) und das Jahr der Veröffentlichung des Datensatzes (2007) das Alter der Versicherten berechnet. Es wird eine neue Variable mit dem Namen „Alter“ generiert.

Im Folgenden werden Fehlerbalkendiagramme über das Alter der Versicherten in den einzelnen Klassen dargestellt. Die Differenz im Alter kann von den Mittelwerten in der Abbildung 3.7 leicht abgelesen werden, wobei die Konfidenzintervalle keine Überlappung aufweisen. Daher kann geschlossen werden, dass sich die Klassen stark anhand des Alters unterscheiden.

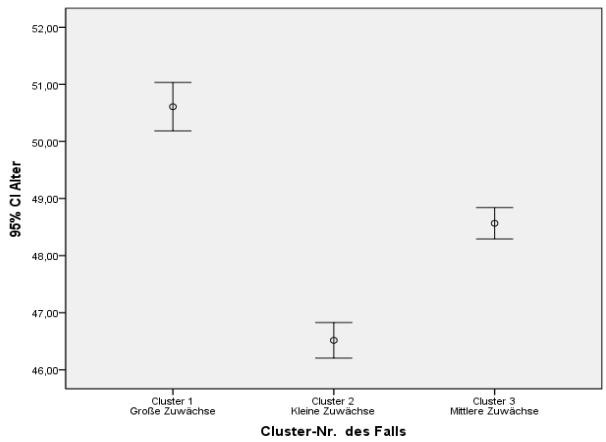


Abbildung 9: *Fehlerbalkendiagramm des Alters in den Clustern*

Folglich kann der Cluster der großen Zuwächse als die Gruppe der älteren Personen bestimmt werden. Im Gegensatz zu dieser Klasse steht der Cluster der kleinen Zuwächse, wo das Durchschnittsalter am niedrigsten ist. Wieder in der Mitte der Ordnungen steht der dritte Cluster, der die mittleren monatlichen Zuwächse der Entgeltpunkte repräsentiert.

Der zweite Faktor, der ins Gewicht fällt, ist das Geschlecht. Im Folgenden wird das Verhältnis (Männer/Frauen) in den einzelnen Clustern abgebildet.

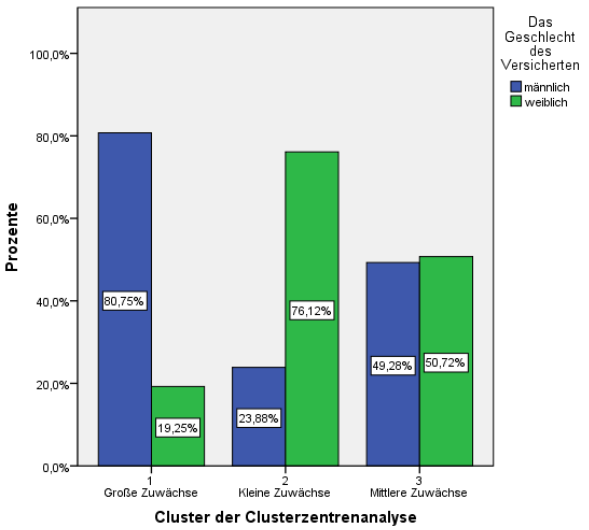


Abbildung 10: *Geschlechterverhältnis in den Clustern*

Vom Balkendiagramm in Abbildung 10 lässt sich nachvollziehen, dass 80,75% der in dem Cluster mit den großen Zuwächsen (Cluster1) enthaltenen Personen männlich sind, während im Cluster mit den niedrigen Zuwächsen die Frauen mit 76,12% überwiegen. Der dritte Cluster, in dem die Versicherten durchschnittliche Zuwächse haben, ist von beiden Geschlechtern fast gleich besetzt.

Nach der Beschreibung der Cluster anhand der soziodemographischen Merkmale werden auch die Werte aus der Rentenberechnung eingeschlossen. Zweifellos wird die Summe der Entgeltpunkte je größer, desto größer die monatlichen Zuwächse und desto älter die Untersuchungspersonen sind. Es ist auch nicht sinnvoll, die Versicherten nach ihren Beitragszeiten zu vergleichen, weil in diesem Fall die älteren Menschen den Vorzug haben würden. Aus diesem Grund fällt der Fokus auf Variablen, die in geringem Maße vom Alter abhängig sind. Vom besonderen Interesse für die Interpretation sind die Merkmale, die die beitragsgeminderten Zeiten und die Anrechnungszeiten repräsentieren. Der Unterschied zwischen den beiden Begriffen liegt in der Festsetzung, ob ein Monat beitragsgemindert oder beitragsfrei ist. Als beitragsfreie Monate oder Anrechnungszeiten werden diese Monate bezeichnet, für die die Versicherten keine Beiträge gezahlt haben. Für beitragsgeminderte Monate werden die Zeiten, die teilweise als beitragsfreie Zeiten und teilweise als Beitragszeiten bestimmt werden, bezeichnet. Die durchschnittliche Anzahl an beitragsgeminderten Zeiten und Anrechnungszeiten sieht folgendermaßen aus (s. Abbildung 11):

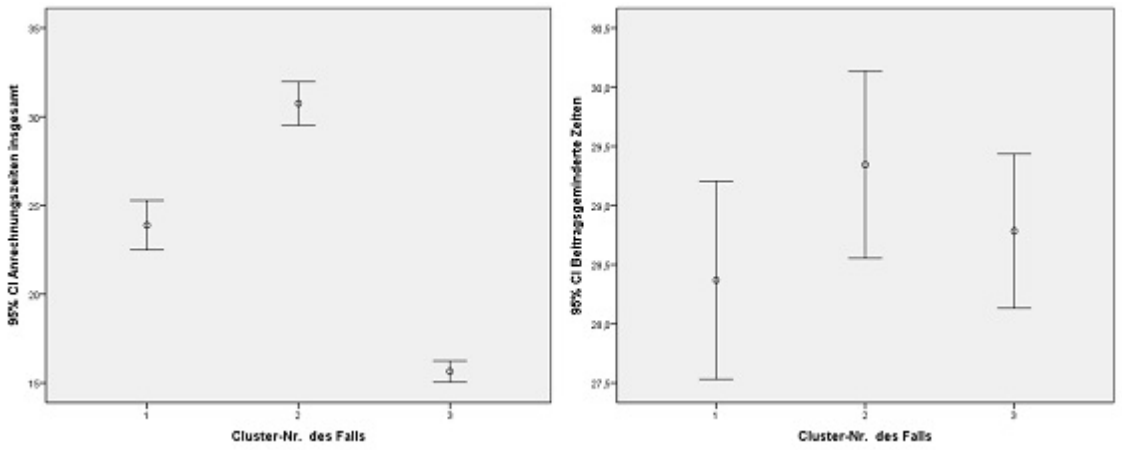


Abbildung 11: *Fehlerbalkendiagramme der Anrechnungszeiten insgesamt (links) und der beitragsgeminderten Zeiten (rechts)*

Im Nächsten werden die Fehlerbalkendiagramme hinsichtlich einiger Faktoren, die grundlegend für die Anrechnungszeiten sind, abgebildet (s. Abb.12). Ein solches Merkmal ist die Variable mit dem Namen „AUAZ“ (Anrechnungszeiten wegen Krankheit). Diese Variable gibt Auskunft über die Anzahl der Monate, in denen der Versicherte als arbeitsunfähig bestimmt wurde. Andere Variable, die für die Interpretation nützlich ist, ist „AJAZ“ (Anrechnungszeiten wegen Arbeitslosigkeit).

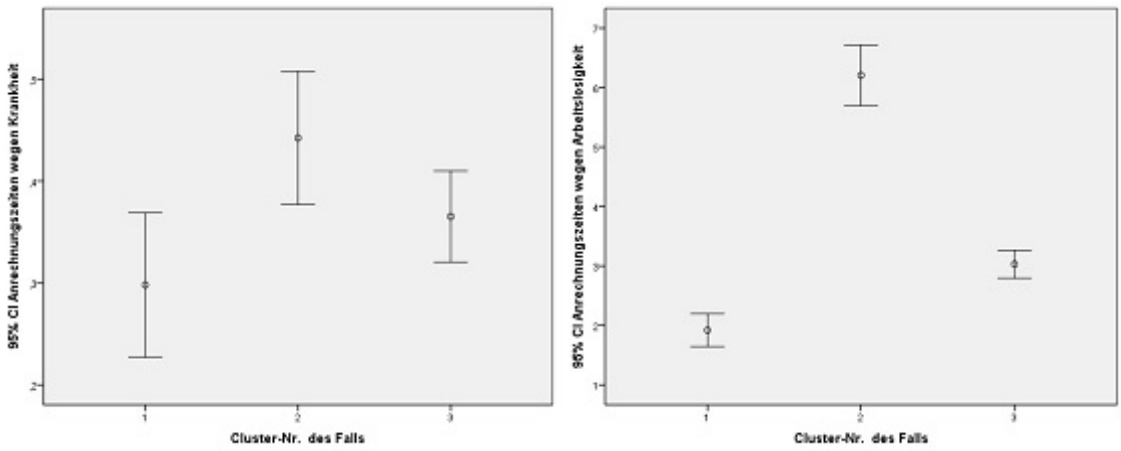


Abbildung 12: *Fehlerbalkendiagramme der Anrechnungszeiten wegen Krankheit (links) und der Anrechnungszeiten wegen Arbeitslosigkeit (rechts)*

Von den Fehlerbalkendiagrammen in Abbildung 12 (links) kann man schlussfolgern, dass die Anrechnungszeiten wegen Krankheit aufgrund der teilweisen Überlappung der Konfidenzintervalle nicht so stark zu der Klassierung der monatlichen Zuwächse beitragen. Dem gegenüber ist vom rechten Diagramm leicht zu erkennen, dass sich die Cluster nach den Anrechnungszeiten wegen Arbeitslosigkeit stark unterscheiden.

Die Variable mit dem Namen „SHULAZ“ (Summe der Anrechnungszeiten wegen schulischer Ausbildung)[11]soll bei der Interpretation besonders berücksichtigt werden, weil sie die Gedanken für die lokale Clusteranalyse entschlüsselt. Da die Variable die Ausbildungsdauer in Monaten repräsentiert, lässt sich leicht berechnen, wie viel Zeit jeder Versicherte in die Ausbildung investiert hat.

[11]Summe der Anrechnungszeiten wegen schulischer Ausbildung fasst alle Anrechnungszeiten in Monaten der Schul- Fachschul- und Hochschulausbildung zusammen

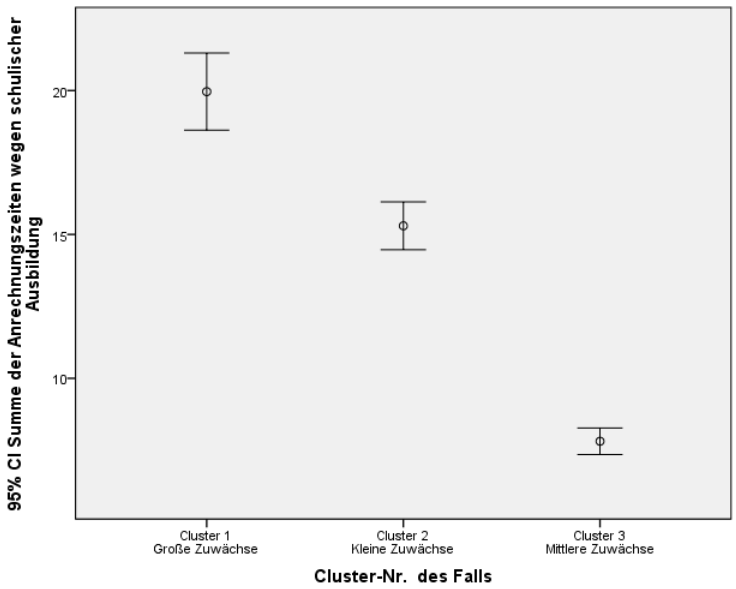


Abbildung 13: *Fehlerbalkendiagramm der Anrechnungszeiten wegen Ausbildung*

Der Unterschied in den Mittelwerten und in ihren Konfidenzintervallen in Abbildung 13 spricht eindeutig für die verschiedenen Ausbildungsniveaus der Versicherten in den einzelnen Clustern. Bemerkenswert ist die Ausbildungsdauer der Personen der Klasse mit den kleinsten Zuwächsen. Obwohl diese Menschen relativ mehr als diejenigen im letzten Cluster gelernt haben, erwerben weniger Entgeltpunkte pro Monat. Diese Feststellung, dass die Frauen in der zweiten Gruppe überwiegen und ihr Durchschnittsalter geringer als dieses der Personen in den anderen Clustern ist, lässt einen Anstieg im Einkommen und einen wahrscheinlichen Wechsel zu einem der beiden Cluster mit größeren Zuwächsen vermuten. Um diese Annahme zu prüfen, ist es nötig, eine zweite Clusteranalyse durchzuführen, wobei konkrete Altersgrenzen zu setzen sind. Dieses Problem wird im nächsten Kapitel ausführlich erläutert.

Nachdem die Cluster bezüglich einiger soziodemographischen Merkmale und Werte aus der Rentenberechnung verglichen werden, lassen sich die folgenden Charakteristiken für die Klassifikation der deutschen Rentenversicherten anhand der Verläufe von Entgeltpunkten ableiten, wobei die Grafiken A.4 und A.5 im Anhang auch als ein Nachweis gelten:

- Cluster 1 beinhaltet die Versicherten, die die größten monatlichen Zuwächse (0,11-0,12) bei der Sammlung von Entgeltpunkten aufweisen. Die Gruppe ist zum großen Teil von Männern besetzt (80,75%), die im Vergleich zu den Personen aus den anderen Clustern ein

höheres Durchschnittsalter haben. Obwohl die Anrechnungszeiten insgesamt für diese Gruppe relativ viel sind, wird von den konkretisierten Fehlerbalkendiagrammen deutlich, dass die Anzahl der beitragsfreien Monate auf die Ausbildung beruht. Die Variable „Anrechnungszeiten wegen Schulausbildung“ ist die einzige, bei der die Personen von Cluster 1 einen größeren Mittelwert im Vergleich zu den anderen Cluster haben.

- Cluster 2 ist von den Versicherten mit den kleinsten Zuwächsen belegt. Die typische Person aus dieser Klasse ist weiblich und hat das niedrigste Durchschnittsalter gegenüber den Untersuchungspersonen aus den anderen Clustern. Von den dargestellten Grafiken lässt sich erkennen, dass diese Person am meisten Anrechnungszeiten hat. Die Versicherten aus dieser Klasse zeichnen sich dadurch aus, dass sie am wahrscheinlichsten arbeitslos und arbeitsunfähig sind. Diesen Fakten gegenüber steht das Balkendiagramm bezüglich der Anrechnungszeiten wegen Schulausbildung, wo es dokumentiert wurde, dass die Personen aus diesem Cluster über mehr Ausbildung als diese mit den mittleren Zuwächsen verfügen. Aus diesem Grund wird erwartet, dass manche Personen in späteren Zeiträumen zu den anderen Clustern wechseln werden.

- Cluster 3 umfasst Personen, die durchschnittlich verdienen und entsprechend mittlere Zuwächse der Entgeltpunkte haben. Beide Geschlechter sind zu gleichen Teilen vertreten und die Versicherten aus dem Cluster mittelmäßiges Durchschnittsalter haben. Das, was für diese Gruppe hervorgehoben werden kann, ist die kleinste Anzahl von Anrechnungszeiten insgesamt. Hier ist zu beachten, dass die Personen jedoch nicht die wenigsten beitragsfreien Zeiten hinsichtlich der Krankheit und der Arbeitslosigkeit haben. Die kleinste Anzahl an Anrechnungszeiten insgesamt kann sich folglich nur dann ergeben, wenn die Anrechnungszeiten wegen Schulausbildung deutlich weniger im Vergleich zu den anderen Anrechnungszeiten sind. Allgemein kann zusammengefasst werden, dass die Versicherten aus dem Cluster mit den mittleren Zuwächsen einen niedrigen Bildungsgrad haben und wahrscheinlich früher den Arbeitsmarkt eingetreten sind.

Im Folgenden werden 3 Fälle aus jedem Cluster ausgewählt und ihre Verlaufskurven über den Jahren hinweg dargestellt. Für diesen Zweck werden Personen, die schon in Rente sind, berücksichtigt. Das zielt auf die Verfügung der Entgeltpunkte über das gesamte Erwerbsalter.

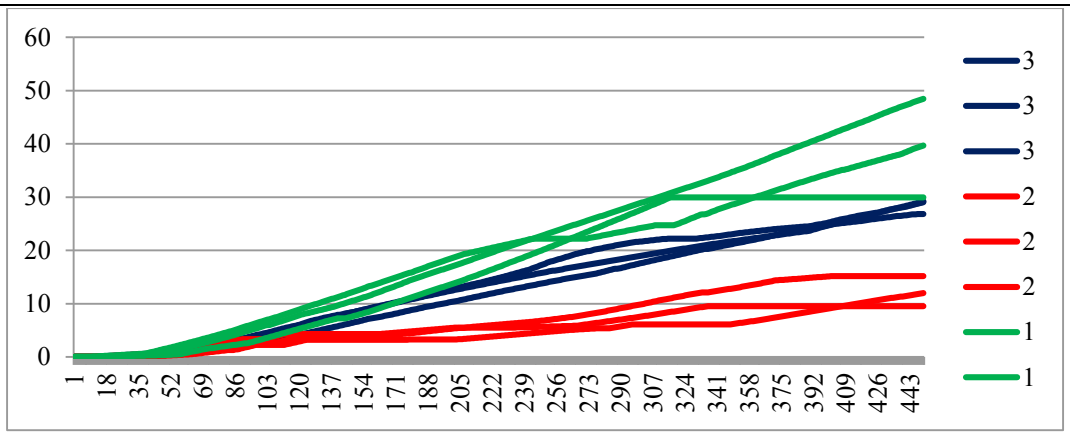


Abbildung 14: *Typische Verläufe von Entgeltpunkten, aufgeteilt nach der Clusterzugehörigkeit*

Anhand der Abbildung 3.12 der drei üblichsten Fälle für jede Gruppe zeigt sich, dass die Verläufe von Entgeltpunkten in drei Cluster gut klassiert sind. Dafür spricht die deutliche Trennung der Verlaufskurven, was auch von der Matrix-Streudiagramme deutlich wird (siehe A.5 im Anhang).

3.4 Lokale Clusteranalyse

Im Abschnitt über die Interpretation der Ergebnisse des globalen Clusterings wurde die Notwendigkeit von einer erneuten Clusteranalyse kurz angesprochen. Die Gedanken für eine neue Klassierung der Versicherten beruhen auf einige Tatsachen, die wir anhand der globalen Clusteranalyse ermittelt haben.

Zuerst ist die Verteilung der Männer und der Frauen zwischen den Clustern zu erwähnen. Von den Ergebnissen wurde festgestellt, dass die Herren größere monatliche Zuwächse als die Damen besitzen. Das ist jedoch keine Neuigkeit, weil auf dem Arbeitsmarkt trotz der Versuche sie zu vermeiden, Diskriminierung bezüglich der Frauen herrscht.[12] Nämlich diese Tatsache weist die Grafik von „The Wall Street Journal“ (s. Abbildung 15) nach. Sie zeigt die Lücke in den Einkommen von Männern und Frauen für die Jahren 1965 bis einschließlich 2010. Man kann leicht erkennen, dass trotz der Verkleinerung dieser Lücke die Männer deutlich mehr als die Frauen verdienen, daher wird es sinnvoll einzelne Clusteranalyse jeweils für die Männer und die Frauen durchzuführen.

[12]Der Zusammenhang zwischen dem Einkommen und den Entgeltpunkten wurde im Kapitel 1.1 dargestellt

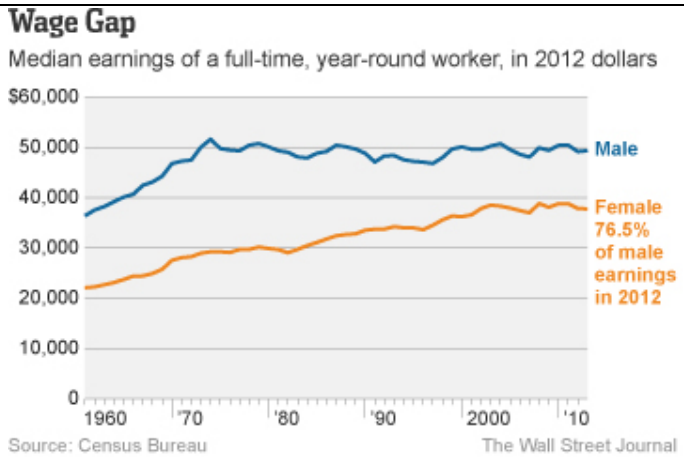


Abbildung 15: *Einkommenslücke zwischen den Geschlechtern*

Anderer Einflussfaktor, der die meisten Frauen in den Cluster mit den kleinen Zuwächsen klassiert, sind die Monate der Schwangerschaft und die Kindererziehungszeiten. Da die Frauen die Hauptlast der Familienpflichten während dieser Zeiten tragen, bleiben sie häufig jahrelang vom Arbeitsmarkt (besonders bei mehr als ein Kind) entfernt. Es werden zwar Entgeltpunkte für diese Monate am Ende des Erwerbsalters zusätzlich berechnet, aber diese Punkte werden von dem Datensatz als monatliche biographiebezogene Merkmale nicht repräsentiert. Dort steht für die Zeiten der Schwangerschaft und Kindererziehung ein Beitrag von Null, falls die Person nicht gearbeitet hat. Diese Tatsache an sich spricht ebenfalls zugunsten eines erneuten Clusterings, aufgeteilt nach Geschlecht.

Von Abbildung 13 lässt sich erkennen, dass die Personen mit den kleinen monatlichen Zuwächsen deutlich mehr in die Ausbildung investiert haben. Aus diesem Grund wurde es hier prognostiziert, dass diese Menschen noch keine „Erträge“ von der Investition bekommen haben. Daher kann man erwarten, dass ein großer Anteil der Versicherten von Cluster 2 zu einem Cluster mit größeren Zuwächsen in den nächsten Zeiträumen übergehen würden.

Damit Einflussfaktoren wie der beitragsfreien Zeiten wegen der Schwangerschaft und der Kindererziehung und der Anrechnungszeiten wegen Ausbildung isoliert werden, soll man eher die Übergänge von einem Cluster zu einem anderen Cluster im Rahmen der weiblichen und männlichen Analyse analysieren. Für diesen Zweck werden einzelne lokale Clusteranalysen für unterschiedliche Perioden im Leben des Rentenversicherten durchgeführt, wobei für die Männer und die Frauen immer jeweils eine einzelne Klassierung zu realisieren ist. Damit dieser Wechsel von einer Gruppe

zu einer anderen dargestellt werden kann, ist es notwendig die wichtigsten für die Formierung der Entgeltpunkte Perioden festzustellen. Es müssen die sogenannten Clustergrenzen gesetzt werden.

3.4.1 Feststellung der Clustergrenzen

Für die männliche lokale Clusteranalyse sind die Altersgrenzen nicht so entscheidend, deswegen kann man sich auf die Bedeutung der Rahmen der Clusteranalyse für die Frauen fokussieren.

Damit die Bedeutung der Mutterschaft eingeschlossen wird, wird das Alter, mit dem die Mütter ihr erstes, zweites und drittes Kind bekommen, berechnet. Für die Ermittlung des Alters kommen die soziodemographischen Merkmale „GBJA“ (Geburtsjahr des Versicherten), „GBKIJ1“(Geburtsjahr des ersten Kindes), „GBKIJ2“(Geburtsjahr des zweiten Kindes) und „GBKIJ3“ (Geburtsjahr des dritten Kindes) zunutze. Als Ergebnis werden die Variablen „Alter_1Kind“, „Alter_2Kind“ und „Alter_3Kind“ generiert, die das Alter der Mütter bei ihrer ersten, zweiten und dritten Geburt angeben.

Zuerst soll man jedoch die Frage beantworten, warum nur die ersten drei potenziellen Schwangerschaften berücksichtigt werden. Von den dargestellten Statistiken in Abbildung 16 erweist sich, dass der Anteil der Frauen mit mehr als drei Kindern vernachlässigbar klein ist.

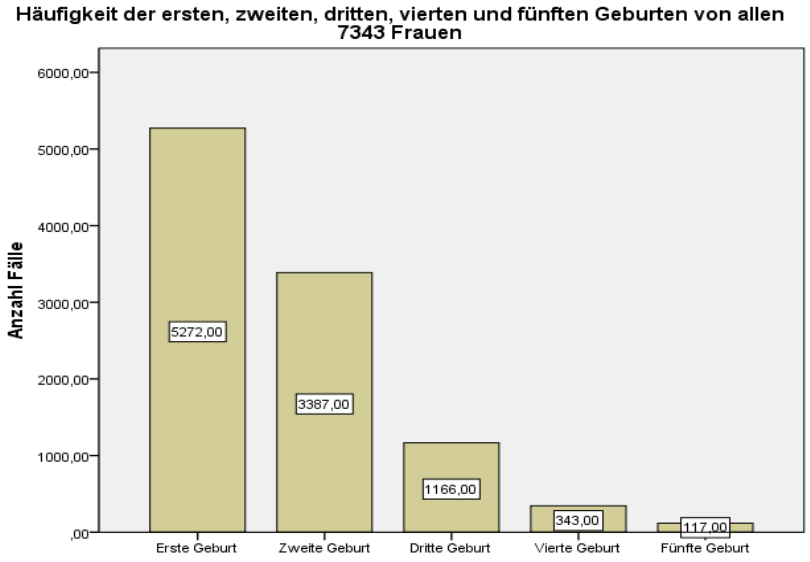


Abbildung 16: *Häufigkeit der ersten, zweiten, dritten, vierten und fünften Geburt*

Die Anzahl von allen 7 343 Frauen, die vier oder mehr Geburten haben, ist 343. Das macht nur 4,7% von den weiblichen Untersuchungspersonen, was wenig Einfluss auf das Clustering hat. Zum Abschluss des Themas der Schwangerschaften ist die deskriptive Statistik des Alters bei den ersten drei Schwangerschaften einzubeziehen.

Deskriptive Statistik

			Statistik	Standardfehler
Alter_1Kind	Mittelwert		24,3994	,06745
	95% Konfidenzintervall des Mittelwertes	Untergrenze	24,2672	
		Obergrenze	24,5316	
	5% getrimmtes Mittel		24,1182	
	Median		23,0000	
	Varianz		24,356	
	Standardabweichung		4,93516	
	Minimum		13,00	
	Maximum		46,00	
	Spannweite		33,00	
	Interquartilbereich		6,00	
	Schiefe		,830	,033
	Kurtosis		,498	,067
Alter_2Kind	Mittelwert		26,9591	,07945
	95% Konfidenzintervall des Mittelwertes	Untergrenze	26,8033	
		Obergrenze	27,1149	
	5% getrimmtes Mittel		26,7903	
	Median		27,0000	
	Varianz		21,610	
	Standardabweichung		4,64862	
	Minimum		15,00	
	Maximum		49,00	
	Spannweite		34,00	
	Interquartilbereich		7,00	
	Schiefe		,529	,042
	Kurtosis		,130	,084
Alter_3Kind	Mittelwert		28,8637	,13535
	95% Konfidenzintervall des Mittelwertes	Untergrenze	28,5982	
		Obergrenze	29,1293	
	5% getrimmtes Mittel		28,7552	
	Median		28,0000	
	Varianz		21,506	
	Standardabweichung		4,63742	
	Minimum		17,00	
	Maximum		43,00	
	Spannweite		26,00	
	Interquartilbereich		7,00	
	Schiefe		,343	,071
	Kurtosis		-,282	,143

Tabelle 9: *Explorative Datenanalyse des Alters bei der ersten, zweiten und dritten Geburt*

Die Werte in Tabelle 9 sind Ausgangspunkt für die Festsetzung der Clustergrenzen. Das Erwerbsalter nach Angaben der Deutschen Rentenversicherung umfasst alle Jahre vom14. bis einschließlich 66. Lebensjahr der Versicherten. Das sind insgesamt 52 Jahre oder 624 Monate. Für die lokale Clusteranalyse wird dieser Zeitraum auf fünf Perioden aufgeteilt, wobei für jede Periode eine Klassierung durchzuführen ist. Jede der fünf Perioden enthält zehn Jahre mit einer einzigen Ausnahme – der letzte Zeitabschnitt umfasst zwölf Jahre.

Folglich sehen die Perioden so aus:

1. Periode: umfasst die Lebensjahre von 14. bis einschließlich 23. Lebensjahr. Da der Median des Alters, mit dem die Frauen ihr erstes Kind bekommen, 23 beträgt, kann angenommen werden, dass in dieser Periode die Hälfte der Mütter ihr erstes Kind gebären. Es ist außerdem zu erwarten, dass ein großer Anteil der Versicherten für diese zehn Jahre mit ihrer Ausbildung fertig wird.
2. Periode enthält die nächsten zehn Lebensjahre – vom 24. bis einschließlich 33. Lebensjahr. Da die Verteilung der Variablen „Alter_1Kind“, „Alter_2Kind“ und „Alter_3Kind“ relativ symmetrisch ist, kann von den Werten für den Mittelwert, die Median und die Standardabweichung zusammengeschlossen werden, dass die Mehrheit der Mütter ihr erstes, zweites und drittes Kind in Rahmen während dieser Periode bekommen werden.
3. Periode: von dem 34. bis einschließlich 43. Lebensjahr. Da die Oberschranke der Periode fast mit dem Ende der Fertilität zusammenfällt, kann man vermuten, dass nach diesem Zeitabschnitt fast keine Geburten mehr stattfinden.
4. Periode: von dem 44. bis einschließlich 53. Lebensjahr. Die Periode wird als die goldenen Jahre gekennzeichnet, weil die Arbeiter in diesem Zeitabschnitt das höchste Einkommen aufweisen
5. Periode: von dem 54. bis einschließlich 66. Lebensjahr. Nach dem erwarteten Anstieg des Einkommens (dementsprechend auch der Entgeltpunkte) in der vorigen Periode kann man hier eine Beibehaltung auf dem gleichen Niveau oder sogar eine Senkung der monatlichen Zuwächse vorhersagen.

Mit anderen Worten werden die folgenden Verläufe während (s. Abbildung 17) für die Frauen und die Männer während des Erwerbsalters erwartet.

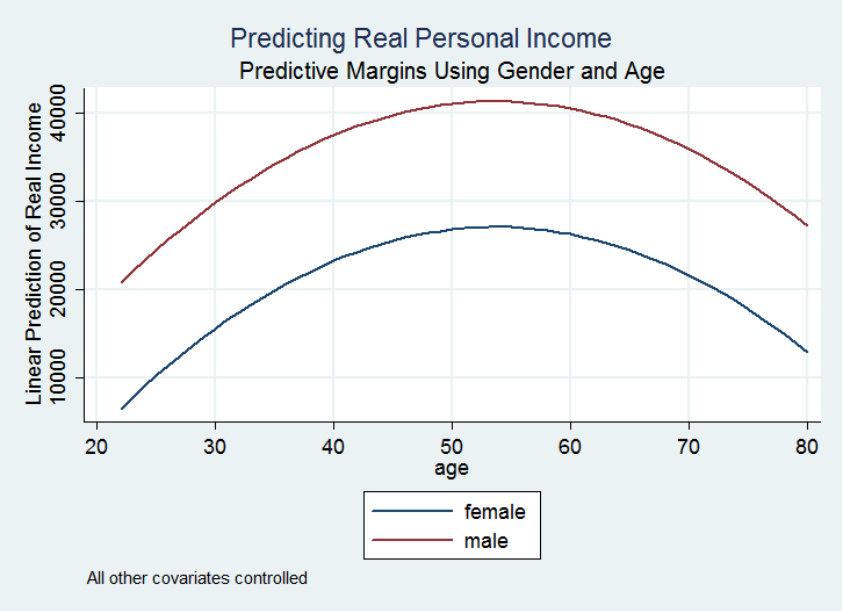


Abbildung 17: *Einfluss des Alters auf das Einkommen*

3.4.2 Die Variablen

Für die lokale Clusteranalyse werden erneut die biographiebezogenen Verlaufsmerkmale, die die erworbenen Entgeltpunkte für jeden Monat repräsentieren, angewendet (s. Tabelle 10). Da im vorigen Kapitel die Clustergrenzen für das neue Clustering festgelegt werden, müssen diesmal die Kennzahlen für den Mittelwert, Median und die Standardabweichung nicht für alle Beitragsmonate errechnet werden, sondern nur für diejenigen, die zu der entsprechenden Periode gehören.[13] Anders ausgedrückt werden zum Beispiel für die erste Periode (14-23) alle Variablen, die die Entgeltpunkte für die ersten 10*12=120 Monate angeben, einbezogen. Das heißt, dass man durch die bekannte Methodik vom Kapitel 3.3.1 den Kalendermonat zu jeder der Variablen von MEGPT_1 bis einschließlich MEGPT_120 bestimmen kann. Am Ende werden die Merkmale in zwölf Gruppen aufgeteilt, wobei jede Gruppe für einen Kalendermonat steht Innerhalb jeder Gruppe werden ihr

[13] Die Beudeutung der drei Kennzahlen für die Clusteranalyse wird ausführlich im Kapitel 3.3.1 erläutert.

Mittelwert, ihr Median und ihre Standardabweichung als neue Variablen berechnet und zum Datensatz addiert.

Variable	Erläuterung
v14_23_*i*MEAN	Mittelwert des i-ten Kalendermonats für Periode 1
v14_23_*i*MEDIAN	Median des i-ten Kalendermonats für Periode 1
v14_23_*i*SD	Standardabweichung des i-ten Kalendermonats für Periode 1
v24_33_*i*MEAN	Mittelwert des i-ten Kalendermonats für Periode 2
v24_33_*i*MEDIAN	Median des i-ten Kalendermonats für Periode 2
v24_33_*i*SD	Standardabweichung des i-ten Kalendermonats für Periode 2
v34_43_*i*MEAN	Mittelwert des i-ten Kalendermonats für Periode 3
v34_43_*i*MEDIAN	Median des i-ten Kalendermonats für Periode 3
v34_43_*i*SD	Standardabweichung des i-ten Kalendermonats für Periode 3
v44_53_*i*MEAN	Mittelwert des i-ten Kalendermonats für Periode 4
v44_53_*i*MEDIAN	Median des i-ten Kalendermonats für Periode 4
v44_53_*i*SD	Standardabweichung des i-ten Kalendermonats für Periode 4
v54_66_*i*MEAN	Mittelwert des i-ten Kalendermonats für Periode 5
v54_66_*i*MEDIAN	Median des i-ten Kalendermonats für Periode 5
v54_66_*i*SD	Standardabweichung des i-ten Kalendarmonats für Periode 5

Tabelle 10: *Variablen der lokalen Clusteranalyse*

3.4.3 Durchführung der lokalen Clusteranalyse

Im Folgenden werden separate lokale Clusteranalysen für die Frauen und für die Männer durchgeführt (vgl. Tabellen 11 bzw. 12). Zuerst soll die Vorgehensweise zur Ermittlung der Clusteranzahl und der monatlichen Zuwächse jeder Klasse kurz erläutert werden.

Für die Feststellung der besten Anzahl der Cluster wird die Two-Step Clusteranalyse angewendet, dabei wird in SPSS die automatische Ermittlung anhand des Akaikes Informationskriteriums ausgewählt. Da für alle lokalen Clusteranalysen das größte Gewicht für die Kategorisierung der Rentenversicherten auf den Medianen und den Mittelwerten liegt, können ihre Werte als die üblichen

Zuwächse der Personen jedes Clusters angenommen werden. Die Analyse zeigt, dass die Verteilung der Zuwächse innerhalb der Cluster symmetrisch und die Mediane und die Mittelwerte keine bzw. vernachlässigbar kleine Abweichungen ergeben. Als Nächstes werden die Anzahl der Gruppen und die üblichen monatlichen Entgeltpunkte für jede Gruppe dargestellt, wobei die Stabilität der errechneten Ergebnisse durch mehrmalige Wiederholung der einzelnen lokalen Clusteranalyse überprüft wird. Aus welchen Gründen, das Two-Step Verfahren zu heterogenen Outputs führen kann, wurde im Kapitel in dem die Methode beschrieben wurde, erläutert

Ergebnisse für die Männer[14]

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Periode 14-23	0,01	0,06		
Periode 24-33	0,04	0,08	0,12	
Periode 34-43	0,05	0,15	0,10	
Periode 44-53	0,03	0,07	0,10	0,16
Periode 54-63	0,01	0,05	0,12	

Tabelle 11: *Ergebnisse der lokalen Clusteranalysen für die Männer*

Ergebnisse für die Frauen

	Cluster 1	Cluster 2
Periode 14-23	0,02	0,06
Periode 24-33	0,03	0,08
Periode 34-43	0,03	0,09
Periode 44-53	0,04	0,12
Periode 54-66	0,01	0,07

Tabelle 12: *Ergebnisse der lokalen Clusteranalysen für die Frauen*

3.4.4 Interpretation der lokalen Clusteranalyse

Die Werte in den Tabellen 11 und 12 deuten darauf hin, dass die lokalen Klassierungen der Rentenversicherten für die einzelnen Zeitabschnitte unterschiedliche Clusteranzahl und unterschiedliche monatliche Zuwächse der Entgeltpunkte für Männer und Frauen liefern.

[14]Die Two-Step Clusteranalyse ergibt für die unterschiedlichen Perioden unterschiedliche Anzahl der Klassen.

Im Folgenden wird der Übergang der Rentenversicherten von einem Cluster in Periode i zu einem anderen Cluster in der Periode $i+1$ anhand der gruppierten Balkendiagrammen abgebildet, wobei auf der X-Achse immer die Prozentänderung und auf der Y-Achse die Clusterzugehörigkeit für Periode i bzw. die entsprechenden Zuwächse stehen. Die Gruppen werden mit Hilfe der Zugehörigkeit zu einer Klasse im Periode $i+1$ aufgeteilt.

Der Wechsel **von der ersten Periode (14-23) in die zweite Periode (24-33)** bestätigt die Hypothese, dass die ersten zehn Jahre sehr wahrscheinlich in der Ausbildung investiert wurden. Dafür sprechen die großen Anstiege der Entgeltpunkte. Das Ergebnis ist plausibel, weil auch die Prognose für die Schwangerschaft und die Kindererziehungszeiten erfüllt ist. Ein Hinweis für diese Behauptung ist das Ergebnis, dass ungefähr die Hälfte der weiblichen Rentenversicherten eine Senkung in den monatlichen Entgeltpunkten aufweist (vgl. Abbildung 18).

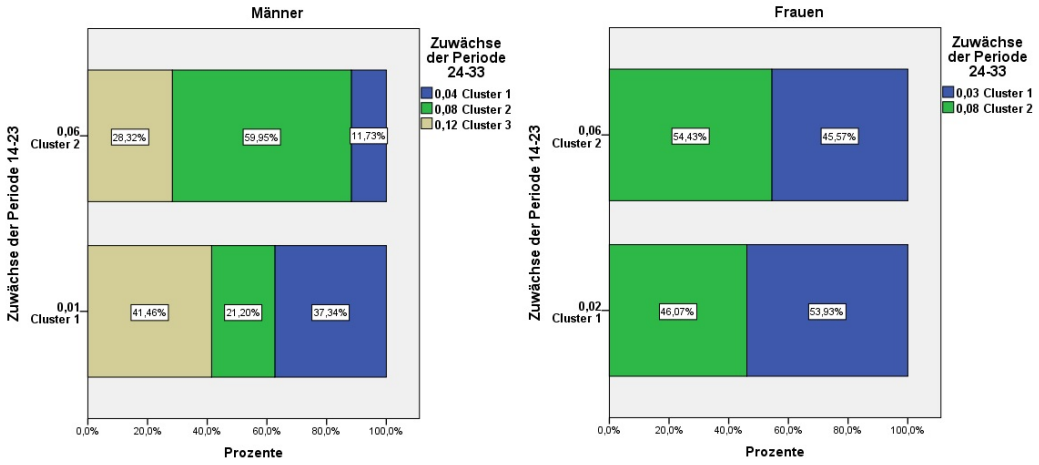


Abbildung 18: *Gestapeltes Balkendiagramm der Wechsel von der ersten Periode in die zweite Periode*

Die Änderungen der Zuwächse **von der zweiten Periode (24-33) in die dritte Periode (34-44)** tragen auch für die Bestätigung der Hypothesen bei. Der Wechsel von einem deutlich größeren Anteil von Männern vom Cluster mit den kleinsten Zuwächsen (etwa 0,04) in den Cluster mit den größten Zuwächsen ist wahrscheinlich auch auf die Ausbildung zurückzuführen. Andererseits weisen die Frauen im Gegensatz zu dem vorigen Diagramm kleinere Abweichungen von dem üblichen Anstieg der Entgeltpunkte aufgrund des Alters. Das bedeutet, dass die meisten Frauen schon die Geburten und die Kindererziehungszeiten hinter sich haben (vgl. Abb. 19).

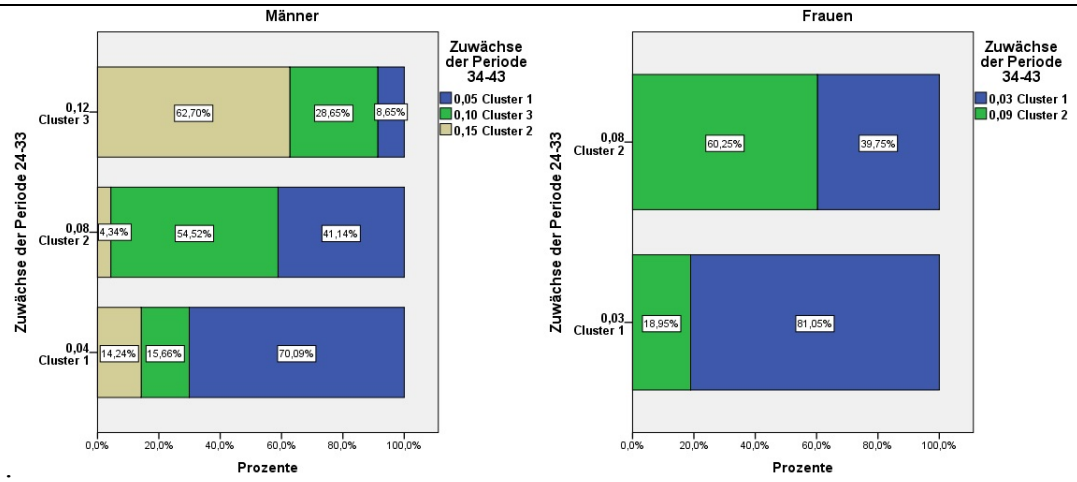


Abbildung 19: *Gestapeltes Balkendiagramm der Wechsel von der zweiten Periode in die dritte Periode*

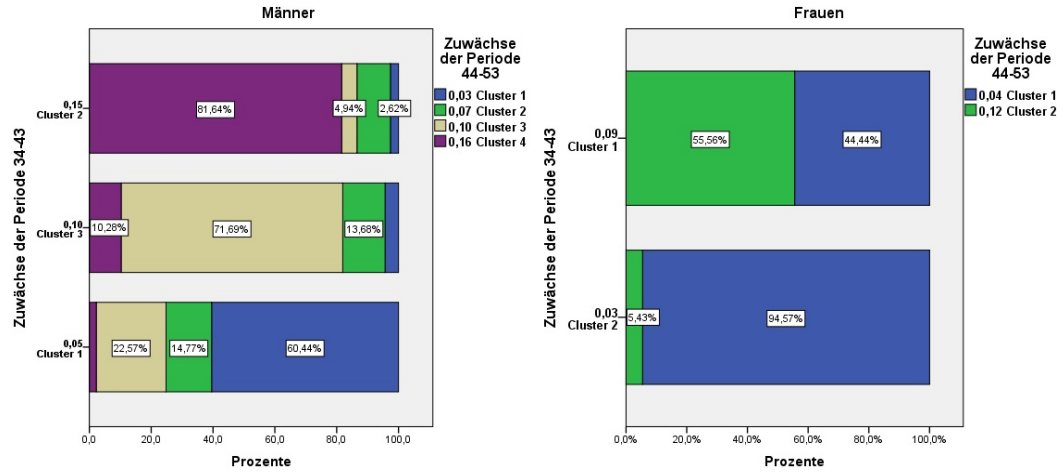


Abbildung 20: *Gestapeltes Balkendiagramm der Wechsel von der dritten in die vierte Periode*

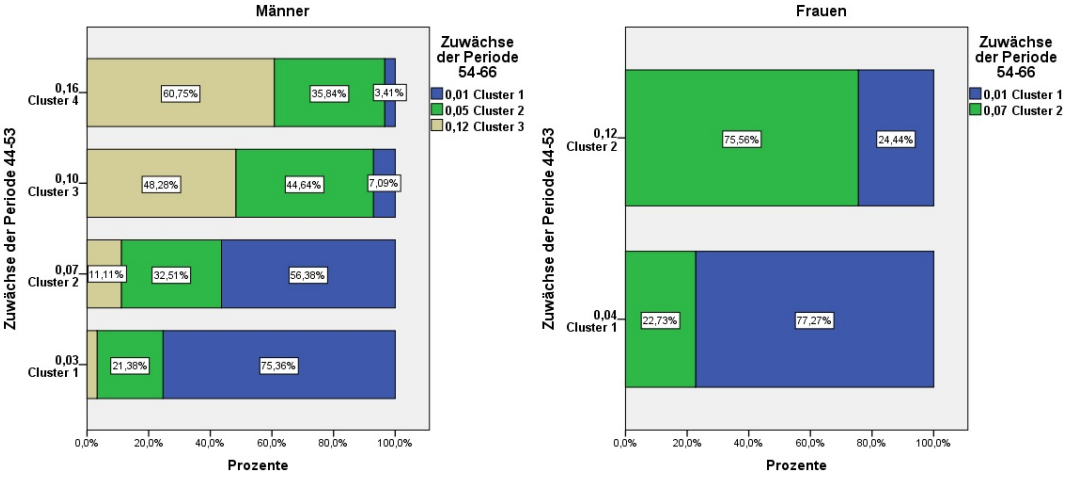


Abbildung 21: *Gestapeltes Balkendiagramm der Wechsel von der vierten in die fünfte Periode*

Die Balkendiagramme (für die dazugehörigen Kreuztabellen siehe den Anhang) der letzten zwei Perioden (s. Abb. 20 bzw. Abb. 21) bestätigen auch die vorigen Feststellungen. Die Männer erwerben, im Vergleich zu den Frauen, mehr Entgeltpunkte im Spitzenalter. Die andere Tatsache, die von den Grafiken abgelesen werden kann, ist, dass die Frauen eine stärkere Prozentänderung von kleinerem Zuwachs in die nächsten Zeitabschnitte aufweisen. Dieses Verhalten der weiblichen Rentenversicherten unterstützt wieder die Hypothese für den Einfluss der Geburten auf dem Verlauf der Entgeltpunkte.

Schließlich kann man zusammenfassen, dass die lokale Clusteranalyse einen näheren Blick auf die wichtigsten Zeiträume im Leben der Versicherten wirft. Die Vermutung über die Bedeutung der Ausbildung, der Ungleichheit im Einkommen und den beitragsfreien Zeiten wegen Schwangerschaft und Kindererziehung für die Klassifikation der deutschen Rentenversicherten wird bestätigt.

4 Zusammenfassung und Schlussfolgerungen

Das Ziel der Arbeit besteht in die Klassifikation der deutschen Rentenversicherten anhand der Verläufe von Entgeltpunkten. Für diesen Zweck wurden die biographiebezogenen Verlaufsmerkmale des Datensatzes „PUFVSKT2007“ eingesetzt, mit deren Hilfe die grundsätzliche Gruppierung der Versicherten durchgeführt wurde.

Das erste Kapitel der Arbeit beschreibt wie sich die Globalisierung, der technologische Fortschritt und der daraus resultierende demographische Wandel auf das Verhalten der Menschen auf dem Arbeitsmarkt auswirken. Kurz erläutert wurden auch die Nebeneinflussfaktoren wie die längere Ausbildungsdauer und der spätere Arbeitsmarkteintritt auf den Versicherungskonten der Untersuchungspersonen. Der Zusammenhang zwischen dem Verhalten auf dem Arbeitsmarkt über die Jahreszeiten hinweg und der Rente wurde mit dem Begriff „Entgeltpunkt“ eingeführt.

Im Kapitel 2 wurden die theoretischen Aspekte der angewandten Clusteranalyse detailliert besprochen. Vorgestellt wurde das ganze Instrumentarium, das man für die Klassifikation der Rentenversicherten anwenden kann. Da das Ähnlichkeits- und Distanzmaß grundlegend für das Clustering ist, wurden auch die üblichsten Maßen für die unterschiedlichen Skalenniveaus vorgestellt.

Im dritten Kapitel der Arbeit wurde die eigentliche Datenanalyse durchgeführt. In der globalen Clusteranalyse wurden die Untersuchungspersonen mittels der Variablen für die monatlichen Mittelwerten, Mediane und Standardabweichungen in einzelnen Clustern gruppiert. Anhand des hierarchischen Verfahrens wurde die Clusteranzahl bei einer Unterstichprobe durchgeführt. Als Nächstes wurden alle Fälle des Datensatzes eingeschlossen und mit Hilfe der Clusterzentrenanalyse gruppiert. Da laut des theoretischen Teils der Arbeit eine Überprüfung der Ergebnisse erforderlich war, wurde auch das Two-Step Verfahren angewandt. Für die Interpretation der Klassifikation wurden soziodemographische Merkmale wie Geschlecht und Alter herangezogen. Mit Hilfe einiger Werte aus der Rentenberechnung wurden Fehlerbalkendiagramme abgebildet, die auch für die Bildung der Profile der drei Cluster beitrugen. Nachdem die Klassen charakterisiert wurden, ließ sich erkennen, dass die älteren Männer die Gruppe mit den großen monatlichen Zuwächsen besetzen, während die jüngeren Frauen eine Zugehörigkeit zu der Klasse mit den kleinen Zuwächsen aufweisen. Die dargestellten Werte aus der Rentenberechnung zeigen, dass der Cluster der Frauen trotz des geringeren Alters am meisten beitragsgeminderte und beitragsfreie Zeiten besitzt. Bei der Betrachtung der Gründe für die Anrechnungszeiten wurde festgestellt, dass die Frauen neben der großen Anzahl an beitragsfreien Monaten wegen Krankheit und Arbeitslosigkeit auch viele beitragsfreie Zeiten wegen Ausbildung haben. Aus diesem Grund war zu erwarten, dass die in Ausbildung investierte Zeit zu größeren Zuwächsen in den späteren Zeiträumen führen würde. Wegen der deutlichen Differenzen zwischen den Klassen im Alter und Geschlecht wurde die Durchführung von einzelnen Clusteranalysen für beide Geschlechter die Klassifikation verfeinern. Damit die erwarteten Änderungen in den monatlichen Zuwächsen aufgrund des Alters und Nebeneinflussfaktoren wie der Schwangerschaft und der Kindererziehungszeit in die Analyse eingeschlossen werden konnten, wurde die Entscheidung getroffen, die lokale Clusteranalyse aufgeteilt nach dem Geschlecht durchzuführen. Dabei wurden die Clustergrenzen so gestaltet, dass jede der gebildeten Perioden durch etwas gekennzeichnet war, was direkt den Verlauf der Entgeltpunkte beeinflusst. Daher wurden die Grenzen der Zeitabschnitte nach dem üblichen Alter für die erste, zweite und dritte Geburt ausgerichtet. Die im globalen Clustering verwendeten Variablen wurden

erneut für jede einzelne Periode berechnet. Die ausgegebenen Cluster und ihre monatlichen Zuwächse bestätigten die formulierte Hypothese, dass Faktoren wie die Ausbildung und die Kindererziehungszeiten eine Wirkung auf die Verläufe der Entgeltpunkte und die entsprechende Klassifikation ausüben.

Schließlich kann man zusammenfassen, dass die Klassifikation der deutschen Rentenversicherten anhand der Verläufe von Entgeltpunkten sich stark für die Männer und die Frauen unterscheidet. Die Abweichungen im Verhalten der Verlaufskurven sind einerseits auf die Ungleichheit im Einkommen und andererseits auch auf die Differenzen in den Familienpflichten zurückzuführen. Diese Annahme kann anhand der Ergebnisse der durchgeführten globalen und lokalen Clusteranalyse weitestgehend bestätigt werden.

Das, was in der vorliegenden Arbeit jedoch nicht analysiert wurde, und von Interesse für weitere Analysen sein könnte, ist die Realisierung einer Faktorenanalyse mittels der Variablen der Entgeltpunkten. Mit dieser Methode können eventuelle latente Faktoren ermittelt werden, die die monatlichen Zuwächse über die Jahre hinweg beeinflussen. Nach der Interpretation und Ekstraktion der Faktoren kann mit ihnen noch eine weitere Klassierung durchgeführt werden.

LITERATURVERZEICHNIS

[1] Brosius, Felix: *SPSS 16: Das mitp-Standardwerk*, *Fundierte Einführung in SPSS und die Statistik*, mitp, 1. Auflage, 2008

[2] Bacher, Johann; Pöge, Andreas; Wenzig, Knüt: *Clusteranalyse, Anwendungsorientierte Einführung in Klassifikationsverfahren*, Oldenbourg Verlag München, 3., ergänzte, vollständig überarbeitete und neu gestaltete Auflage, München, 2010

[3] Bühl, Achim; Zöfel, Peter: *SPSS Version 10, Einführung in die moderne Datenanalyse unter Windows*, ADISSON-WESLEY, 7., überarbeitete und erweiterte Auflage, München, 2000

[4] Bühl, Achim: SPSS 16, *Einführung in die moderne Datenanalyse*, Pearson Studium, 11. Auflage, München, 2008

[5] Bortz, Jürgen: *Statistik für Human- und Sozialwissenschaftler*, Springer, 6. Auflage, Berlin, 2004

[6] The Wall Street Journal; URL: http://blogs.wsj.com/economics/2013/09/17/male-female-pay-gap-hasnt-moved-much-in-years/

[7] Deutsche Rentenversicherung, Lexikon; URL: http://www.deutsche-rentenversicherung.de/Allgemein/de/Navigation/5_Services/01_kontakt_und_beratung/02_beratung/05_lexikon_node.html

ANHANG

A.1 *Anfängliche Clusterzentren, die mittels der hierarchische Clusteranalyse ermittelt wurden*

Anfängliche Clusterzentren

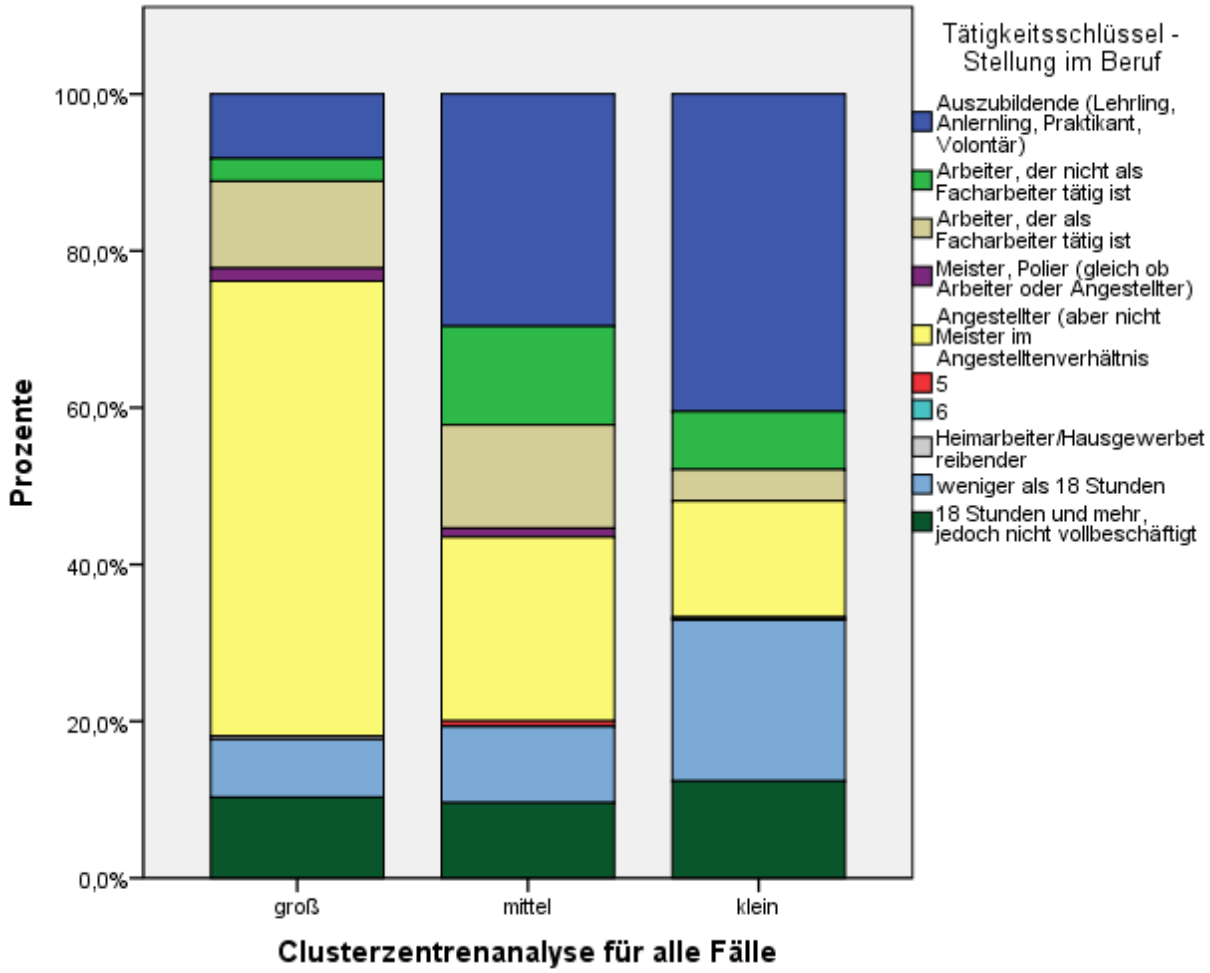
	Cluster		
	1	2	3
MEAN_JANUAR	,10	,03	,06
MEDIAN_JANUAR	,12	,03	,07
SD_JANUAR	,05	,03	,03
MEAN_FEBRUAR	,09	,03	,06
MEDIAN_FEBRUAR	,11	,02	,06
SD_FEBRUAR	,04	,02	,03
MEAN_MÄRZ	,10	,03	,06
MEDIAN_MÄRZ	,12	,03	,07
SD_MÄRZ	,04	,03	,03
MEAN_APRIL	,10	,03	,06
MEDIAN_APRIL	,12	,03	,07
SD_APRIL	,04	,02	,03
MEAN_MAI	,10	,03	,06
MEDIAN_MAI	,12	,03	,07
SD_MAI	,05	,03	,03
MEAN_JUNI	,10	,03	,06
MEDIAN_JUNI	,11	,03	,07
SD_JUNI	,04	,02	,03
MEAN_JULI	,10	,03	,06
MEDIAN_JULI	,12	,03	,07
SD_JULI	,05	,03	,03
MEAN_AUGUST	,10	,03	,06
MEDIAN_AUGUST	,12	,03	,07
SD_AUGUST	,05	,03	,03
MEAN_SEPTEMBER	,10	,03	,06
MEDIAN_SEPTEMBER	,11	,03	,07
SD_SEPTEMBER	,04	,02	,03
MEAN_OKTOBER	,10	,03	,06
MEDIAN_OKTOBER	,12	,03	,07
SD_OKTOBER	,05	,03	,03
MEAN_NOVEMBER	,10	,03	,06
MEDIAN_NOVEMBER	,11	,03	,07
SD_NOVEMBER	,04	,02	,03
MEAN_DEZEMBER	,10	,03	,06
MEDIAN_DEZEMBER	,12	,02	,07
SD_DEZEMBER	,05	,03	,03

A.2 *Clusterzentren der endgültigen Lösung, die mittels der Clusterzentrenanalyse ermittelt wurden*

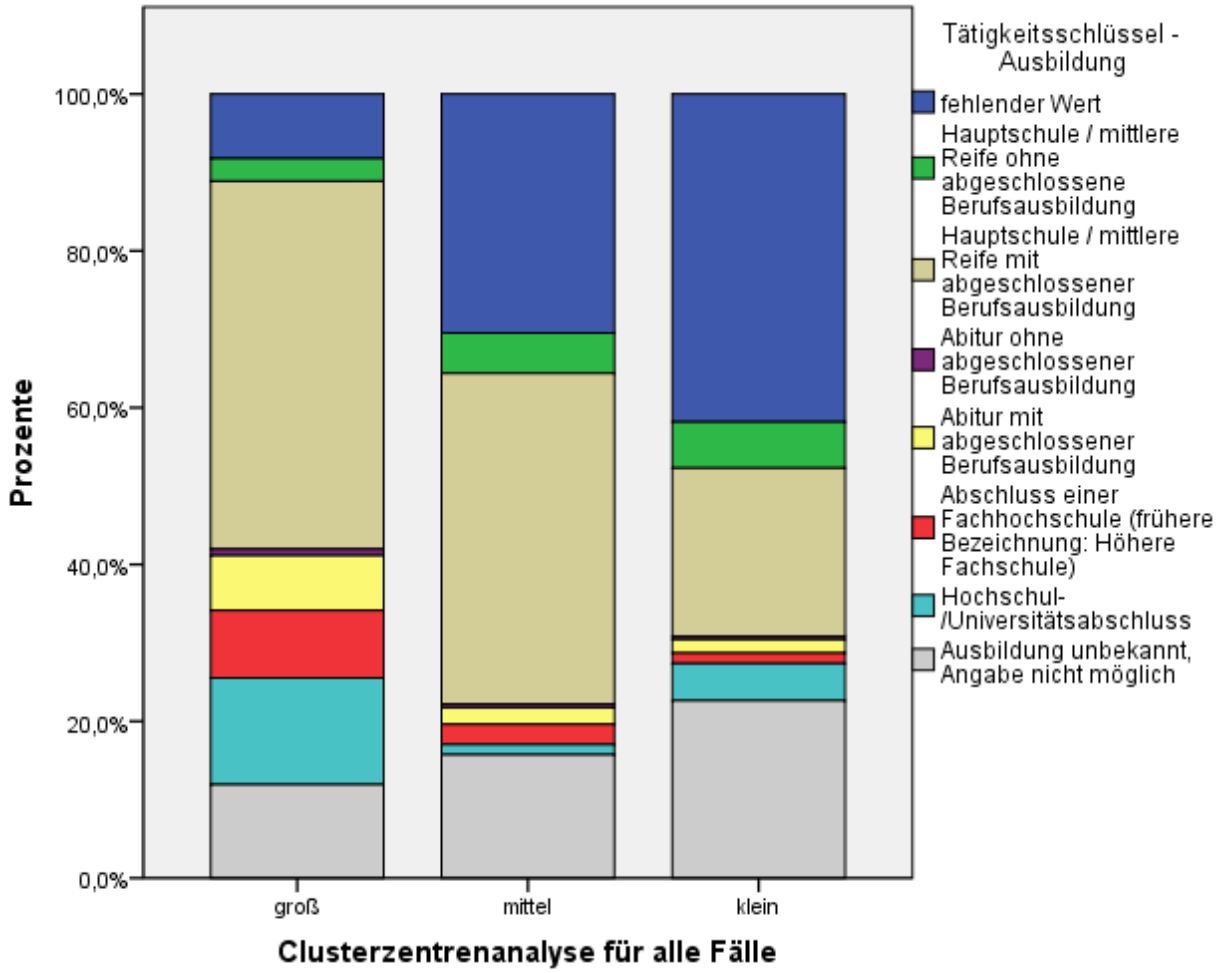
Clusterzentren der endgültigen Lösung

	Cluster		
	1	2	3
MEAN_JANUAR	,10	,03	,06
MEDIAN_JANUAR	,12	,03	,07
SD_JANUAR	,05	,03	,03
MEAN_FEBRUAR	,09	,03	,06
MEDIAN_FEBRUAR	,11	,03	,06
SD_FEBRUAR	,04	,02	,03
MEAN_MÄRZ	,10	,03	,06
MEDIAN_MÄRZ	,12	,03	,07
SD_MÄRZ	,05	,03	,03
MEAN_APRIL	,10	,03	,06
MEDIAN_APRIL	,11	,03	,07
SD_APRIL	,04	,02	,03
MEAN_MAI	,10	,03	,06
MEDIAN_MAI	,12	,03	,07
SD_MAI	,05	,03	,03
MEAN_JUNI	,10	,03	,06
MEDIAN_JUNI	,11	,03	,07
SD_JUNI	,04	,02	,03
MEAN_JULI	,10	,03	,06
MEDIAN_JULI	,12	,03	,07
SD_JULI	,05	,03	,03
MEAN_AUGUST	,10	,03	,06
MEDIAN_AUGUST	,12	,03	,07
SD_AUGUST	,05	,03	,03
MEAN_SEPTEMBER	,10	,03	,06
MEDIAN_SEPTEMBER	,11	,03	,07
SD_SEPTEMBER	,05	,02	,03
MEAN_OKTOBER	,10	,03	,06
MEDIAN_OKTOBER	,12	,03	,07
SD_OKTOBER	,05	,03	,03
MEAN_NOVEMBER	,10	,03	,06
MEDIAN_NOVEMBER	,11	,03	,07
SD_NOVEMBER	,05	,02	,03
MEAN_DEZEMBER	,10	,03	,06
MEDIAN_DEZEMBER	,12	,02	,07
SD_DEZEMBER	,05	,03	,03

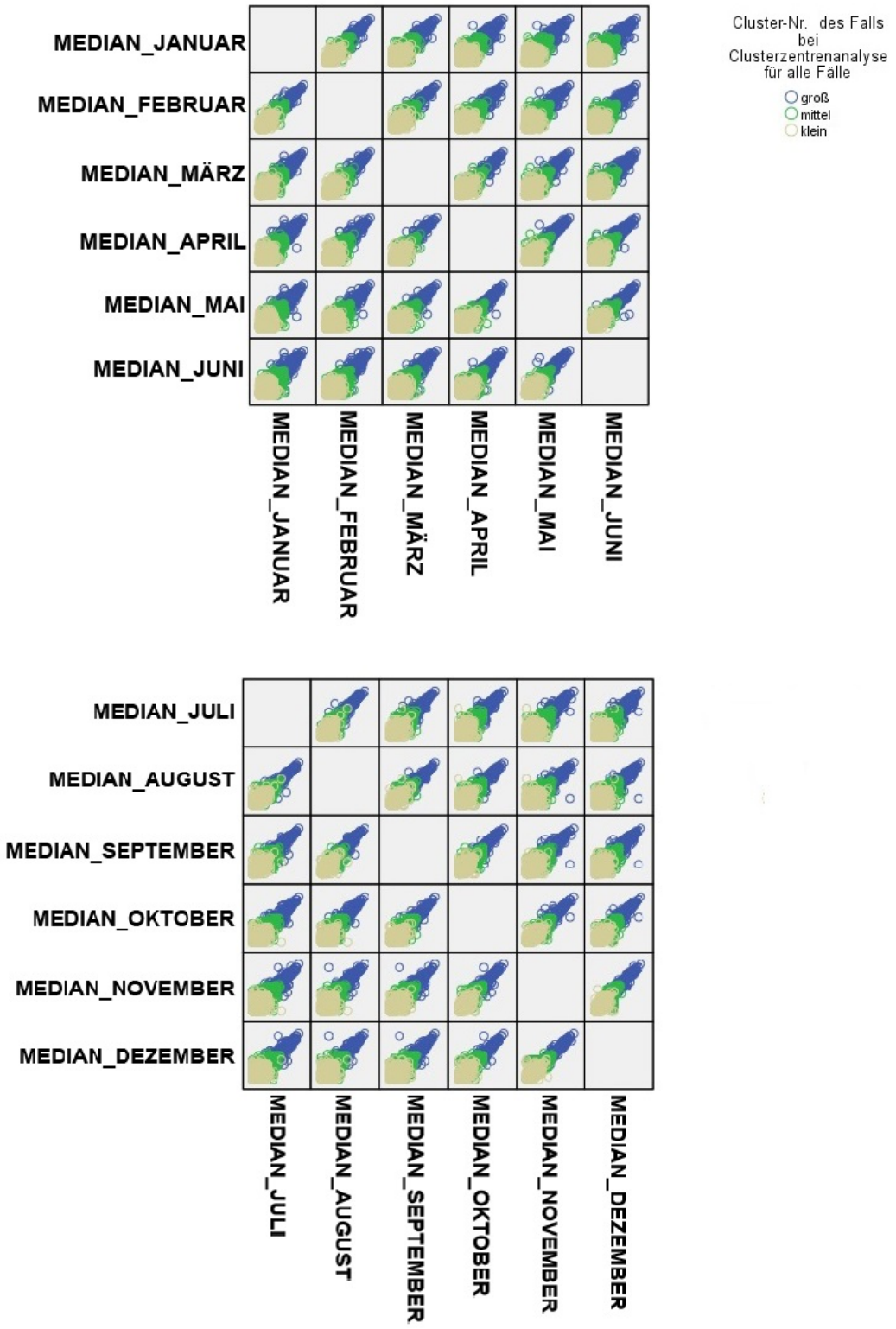
A.3 *Tätigkeitsschlüssel – Stellung im Beruf innerhalb der Cluster*



A.4 *Tätigkeitsschlüssel – Ausbildung innerhalb der Cluster*



A.5 *Matrix-Streuungsdiagramm*



A.6 *Wechsel der Männer von der ersten in die zweite Periode*

Kreuztabelle

			Monatliche Zuwächse der zweiten Periode			
			0,04	0,08	0,12	Gesamt
Monatliche Zuwächse der ersten Periode	0,01	Anzahl	481	273	534	1288
		Erwartete Anzahl	233,5	647,4	407,1	1288,0
		% zeilenweise	37,3%	21,2%	41,5%	100,0%
		% spaltenweise	51,5%	10,5%	32,8%	25,0%
		% der Gesamtzahl	9,3%	5,3%	10,4%	25,0%
	0,06	Anzahl	453	2316	1094	3863
		Erwartete Anzahl	700,5	1941,6	1220,9	3863,0
		% zeilenweise	11,7%	60,0%	28,3%	100,0%
		% spaltenweise	48,5%	89,5%	67,2%	75,0%
		% der Gesamtzahl	8,8%	45,0%	21,2%	75,0%
Gesamt		Anzahl	934	2589	1628	5151
		Erwartete Anzahl	934,0	2589,0	1628,0	5151,0
		% zeilenweise	18,1%	50,3%	31,6%	100,0%
		% spaltenweise	100,0%	100,0%	100,0%	100,0%
		% der Gesamtzahl	18,1%	50,3%	31,6%	100,0%

A.7 *Wechsel der Männer von der zweiten in die dritte Periode*

Kreuztabelle

			Monatliche Zuwächse der dritten Periode			
			0,05	0,15	0,10	Gesamt
Monatliche Zuwächse der zweiten Periode	0,04	Anzahl	443	90	99	632
		Erwartete Anzahl	219,1	159,0	253,9	632,0
		% zeilenweise	70,1%	14,2%	15,7%	100,0%
		% spaltenweise	30,0%	8,4%	5,8%	14,9%
		% der Gesamtzahl	10,4%	2,1%	2,3%	14,9%
	0,08	Anzahl	910	96	1206	2212
		Erwartete Anzahl	767,0	556,4	888,6	2212,0
		% zeilenweise	41,1%	4,3%	54,5%	100,0%
		% spaltenweise	61,7%	9,0%	70,6%	52,0%
		% der Gesamtzahl	21,4%	2,3%	28,3%	52,0%
	0,12	Anzahl	122	884	404	1410
		Erwartete Anzahl	488,9	354,7	566,5	1410,0
		% zeilenweise	8,7%	62,7%	28,7%	100,0%
		% spaltenweise	8,3%	82,6%	23,6%	33,1%
		% der Gesamtzahl	2,9%	20,8%	9,5%	33,1%
Gesamt		Anzahl	1475	1070	1709	4254
		Erwartete Anzahl	1475,0	1070,0	1709,0	4254,0
		% zeilenweise	34,7%	25,2%	40,2%	100,0%
		% spaltenweise	100,0%	100,0%	100,0%	100,0%
		% der Gesamtzahl	34,7%	25,2%	40,2%	100,0%

A.8 *Wechsel der Männer von der dritten in die vierten Periode*

Kreuztabelle

			Monatliche Zuwächse der vierten Periode				
			0,03	0,07	0,10	0,16	Gesamt
Monatliche Zuwächse der dritten Periode	0,05	Anzahl	573	140	214	21	948
		Erwartete Anzahl	221,1	126,8	369,3	230,8	948,0
		% zeilenweise	60,4%	14,8%	22,6%	2,2%	100,0%
		% spaltenweise	89,5%	38,1%	20,0%	3,1%	34,5%
		% der Gesamtzahl	20,9%	5,1%	7,8%	,8%	34,5%
	0,15	Anzahl	17	70	32	529	648
		Erwartete Anzahl	151,1	86,7	252,4	157,7	648,0
		% zeilenweise	2,6%	10,8%	4,9%	81,6%	100,0%
		% spaltenweise	2,7%	19,1%	3,0%	79,2%	23,6%
		% der Gesamtzahl	,6%	2,6%	1,2%	19,3%	23,6%
	0,10	Anzahl	50	157	823	118	1148
		Erwartete Anzahl	267,8	153,5	447,2	279,5	1148,0
		% zeilenweise	4,4%	13,7%	71,7%	10,3%	100,0%
		% spaltenweise	7,8%	42,8%	77,0%	17,7%	41,8%
		% der Gesamtzahl	1,8%	5,7%	30,0%	4,3%	41,8%
Gesamt		Anzahl	640	367	1069	668	2744
		Erwartete Anzahl	640,0	367,0	1069,0	668,0	2744,0
		% zeilenweise	23,3%	13,4%	39,0%	24,3%	100,0%
		% spaltenweise	100,0%	100,0%	100,0%	100,0%	100,0%
		% der Gesamtzahl	23,3%	13,4%	39,0%	24,3%	100,0%

A.9 *Wechsel der Männer von der vierten in die fünften Periode*

Kreuztabelle

			Monatliche Zuwächse der fünften Periode			
			0,01	0,05	0,12	Gesamt
Monatliche Zuwächse der vierten Periode	0,03	Anzahl	208	59	9	276
		Erwartete Anzahl	81,1	98,5	96,4	276,0
		% zeilenweise	75,4%	21,4%	3,3%	100,0%
		% spaltenweise	53,1%	12,4%	1,9%	20,7%
		% der Gesamtzahl	15,6%	4,4%	,7%	20,7%
	0,07	Anzahl	137	79	27	243
		Erwartete Anzahl	71,4	86,7	84,9	243,0
		% zeilenweise	56,4%	32,5%	11,1%	100,0%
		% spaltenweise	34,9%	16,6%	5,8%	18,2%
		% der Gesamtzahl	10,3%	5,9%	2,0%	18,2%
	0,10	Anzahl	37	233	252	522
		Erwartete Anzahl	153,4	186,3	182,3	522,0
		% zeilenweise	7,1%	44,6%	48,3%	100,0%
		% spaltenweise	9,4%	48,9%	54,1%	39,1%
		% der Gesamtzahl	2,8%	17,5%	18,9%	39,1%
	0,16	Anzahl	10	105	178	293
		Erwartete Anzahl	86,1	104,5	102,4	293,0
		% zeilenweise	3,4%	35,8%	60,8%	100,0%
		% spaltenweise	2,6%	22,1%	38,2%	22,0%
		% der Gesamtzahl	,7%	7,9%	13,3%	22,0%
Gesamt		Anzahl	392	476	466	1334
		Erwartete Anzahl	392,0	476,0	466,0	1334,0
		% zeilenweise	29,4%	35,7%	34,9%	100,0%
		% spaltenweise	100,0%	100,0%	100,0%	100,0%
		% der Gesamtzahl	29,4%	35,7%	34,9%	100,0%

A.10 *Wechsel der Frauen von der ersten in die zweite Periode*

Kreuztabelle

			Monatliche Zuwächse der zweiten Periode		
			0,03	0,08	Gesamt
Monatliche Zuwächse der ersten Periode	0,02	Anzahl	199	170	369
		Erwartete Anzahl	181,5	187,5	369,0
		% zeilenweise	53,9%	46,1%	100,0%
		% spaltenweise	47,4%	39,2%	43,2%
		% der Gesamtzahl	23,3%	19,9%	43,2%
	0,06	Anzahl	221	264	485
		Erwartete Anzahl	238,5	246,5	485,0
		% zeilenweise	45,6%	54,4%	100,0%
		% spaltenweise	52,6%	60,8%	56,8%
		% der Gesamtzahl	25,9%	30,9%	56,8%
Gesamt		Anzahl	420	434	854
		Erwartete Anzahl	420,0	434,0	854,0
		% zeilenweise	49,2%	50,8%	100,0%
		% spaltenweise	100,0%	100,0%	100,0%
		% der Gesamtzahl	49,2%	50,8%	100,0%

A.11 *Wechsel der Frauen von der zweiten in die dritte Periode*

Kreuztabelle

			Monatliche Zuwächse der ersten Periode		
			0,03	0,09	Gesamt
Monatliche Zuwächse der zweiten Periode	0,03	Anzahl	278	65	343
		Erwartete Anzahl	209,4	133,6	343,0
		% zeilenweise	81,0%	19,0%	100,0%
		% spaltenweise	68,5%	25,1%	51,6%
		% der Gesamtzahl	41,8%	9,8%	51,6%
	0,08	Anzahl	128	194	322
		Erwartete Anzahl	196,6	125,4	322,0
		% zeilenweise	39,8%	60,2%	100,0%
		% spaltenweise	31,5%	74,9%	48,4%
		% der Gesamtzahl	19,2%	29,2%	48,4%
Gesamt		Anzahl	406	259	665
		Erwartete Anzahl	406,0	259,0	665,0
		% zeilenweise	61,1%	38,9%	100,0%
		% spaltenweises	100,0%	100,0%	100,0%
		% der Gesamtzahl	61,1%	38,9%	100,0%

A.12 *Wechsel der Frauen von der dritten in die vierte Periode*

Kreuztabelle

			Monatliche Zuwächse der vierten Periode		
			0,04	0,12	Gesamt
Monatliche Zuwächse der dritten Periode	0,03	Anzahl	244	14	258
		Erwartete Anzahl	195,9	62,1	258,0
		% zeilenweise	94,6%	5,4%	100,0%
		% spaltenweise	78,2%	14,1%	62,8%
		% der Gesamtzahl	59,4%	3,4%	62,8%
	0,09	Anzahl	68	85	153
		Erwartete Anzahl	116,1	36,9	153,0
		% zeilenweise	44,4%	55,6%	100,0%
		% spaltenweise	21,8%	85,9%	37,2%
		% der Gesamtzahl	16,5%	20,7%	37,2%
Gesamt		Anzahl	312	99	411
		Erwartete Anzahl	312,0	99,0	411,0
		% zeilenweise	75,9%	24,1%	100,0%
		% spaltenweise	100,0%	100,0%	100,0%
		% der Gesamtzahl	75,9%	24,1%	100,0%

A.13 *Wechsel der Frauen von der vierten in die fünfte Periode*

Kreuztabelle

			Monatliche Zuwächse der fünften Periode		
			0,01	0,07	Gesamt
Monatliche Zuwächse der vierten Periode	0,04	Anzahl	119	35	154
		Erwartete Anzahl	100,6	53,4	154,0
		% zeilenweise	77,3%	22,7%	100,0%
		% spaltenweise	91,5%	50,7%	77,4%
		% der Gesamtzahl	59,8%	17,6%	77,4%
	0,12	Anzahl	11	34	45
		Erwartete Anzahl	29,4	15,6	45,0
		% zeilenweise	24,4%	75,6%	100,0%
		% spaltenweise	8,5%	49,3%	22,6%
		% der Gesamtzahl	5,5%	17,1%	22,6%
Gesamt		Anzahl	130	69	199
		Erwartete Anzahl	130,0	69,0	199,0
		% zeilenweise	65,3%	34,7%	100,0%
		% spaltenweise	100,0%	100,0%	100,0%
		% der Gesamtzahl	65,3%	34,7%	100,0%

Erklärung zur Urheberschaft

Hiermit erkläre ich, Ivan Mitkov, dass ich die vorliegende Arbeit allein und nur unter Verwendung der aufgeführten Quellen und Hilfsmittel angefertigt habe.
Die Prüfungsordnung ist mir bekannt. Ich habe in meinem Studienfach bisher keine Bachelorarbeit eingereicht bzw. diese nicht endgültig nicht bestanden.

.....
/Mitkov, Ivan/
Berlin, den 24. März 2014